MCMC Using Hamiltonian Dynamics

Maryam Eskandari

October 10, 2022

Introduction

- Markov chain Monte Carlo (MCMC) originated with the classic paper of Metropolis et al (1953).
  - To simulate the distribution of states for a system of idealized molecules
- Hamiltonian Monte Carlo (HMC): molecular simulation was introduced (Alder and Wainwright, 1959), in which the motion of the molecules was deterministic.

These approaches are asymptotically equivalent.

Applications

- Molecular simulation
- Lattice field theory simulations of quantum chromodynamics
- Neural network models

Elements of HMC

- Define a Hamiltonian function in terms of the probability distribution we wish to sample from.
- Position variables: the variables we are interested in
- Momentum variables: typically have independent Gaussian distributions.
- Simple updates using Metropolis updates.
- A new state is proposed by computing a trajectory according to Hamiltonian dynamics.
- The new state will have a high probability of acceptance.
- This bypasses the slow exploration of the state space that occurs when Metropolis updates are done using a simple random-walk proposal distribution

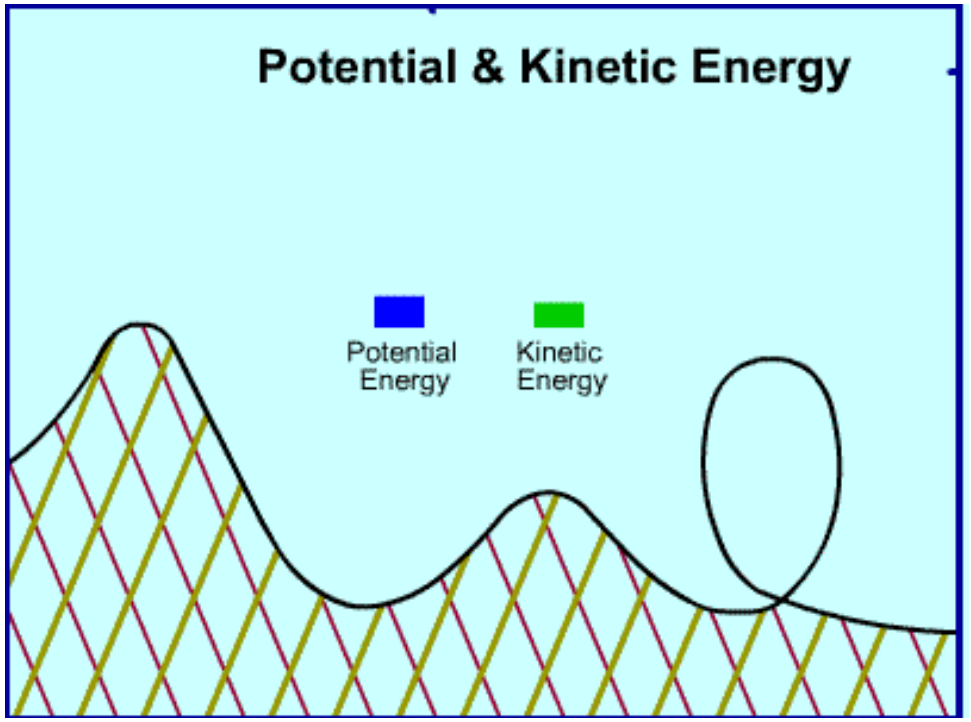
Hamiltonian Dynamics

- Has a physical interpretation
- Like puck that slides over a surface of varying height

The system consists of:

  - Position of the puck given by two-dimensional vector q in 2D space
  - Momentum of the puck given by a two-dimensional vector p.

- The potential energy, U(q), of the puck is proportional to the height of the surface at its current position.
- Kinetic energy, K(p), is equal to $|p|^2/(2m)$, where m is the mass of the puck.



Nonphysical MCMC applications of Hamiltonian dynamics

- The position will correspond to the variables of interest
- The potential energy will be minus the log of the probability density for these variables
- Momentum variables, one for each position variable, will be introduced artificially

Hamilton's Equations

- Position vector q
- Momentum vector p
- Hamiltonian function $H(q,p)$

Partial derivatives describe motion over time t:

$$\frac{dq_i}{dt} = \frac{\partial H}{\partial p_i} \qquad \frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i}$$

Or we can combine p and q into z=(q,p) with 2D dimensional space:

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

∇H is the gradient of H (i.e. $[\nabla H]_k = \frac{\partial H}{\partial z_k}$) and $j = \begin{bmatrix} 0_{d\times d} & I_{d\times d} \\ -I_{d\times d} & 0_{d\times d} \end{bmatrix}$

Is 2d × 2d matrix defined by identity and zero matrices.

Potential and Kinetic energy

Hamiltonian function can be written as:

$$H(q,p)=U(q)+K(p)$$

U(q): *Potential energy* will be defined to be minus the log probability density of the distribution for q that we wish to sample.

K(p): Kinetic energy, and is usually defined as $K(p) = \frac{p^T M^{-1} p}{2}$

M is mass matrix (positive-definite and symmetric, typically diagonal)

This form for K(p) corresponds to minus the log probability density (plus a constant) of the zero-mean Gaussian distribution with covariance matrix M.

With these Hamiltonian equations can be written as:

$$\frac{dq_i}{dt} = [M^{-1}p]_i \qquad \frac{dp_i}{dt} = -\frac{\partial U}{\partial q_i}$$

A one-dimensional example

- q and p are scalars

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

- This corresponds to a Gaussian distribution for q with mean zero and variance one (will be discussed later)
- dynamics resulting from this Hamiltonian:

$$\frac{dq}{dt} = p, \quad \frac{dp}{dt} = -q.$$

- Solutions (for some constant r and a):

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

Properties of Hamiltonian Dynamics

- Reversibility

$$(q(t), p(t)) \rightarrow (q(t+s), p(t+s)) \text{ using } T_s \text{ mapping}$$

$$(q(t+s), p(t+s)) \rightarrow (q(t), p(t)) \text{ using } T_{-s} \text{ mapping}$$

This inverse mapping is obtained by simply negating the time derivatives in Equations:

$$\frac{dq_i}{dt} = \frac{\partial H}{\partial p_i} \qquad \frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i}$$

In the simple one-dimensional example, T_{-s} is just a counterclockwise rotation by s radians, undoing the clockwise rotation of T_s.

Properties of Hamiltonian Dynamics

- Conservation of the Hamiltonian (keeps the Hamiltonian invariant)

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

Since: $\frac{dq_i}{dt} = \frac{\partial H}{\partial p_i}$ $\frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i}$

For Metropolis updates using a proposal found by Hamiltonian dynamics, which form part of the HMC method, the acceptance probability is one if H is kept invariant.

Properties of Hamiltonian Dynamics

- Volume preservation: Hamiltonian dynamics might stretch a region in one direction, as long as the region is squashed in some other direction so as to preserve volume

- Symplecticness: Volume preservation is also a consequence of Hamiltonian dynamics being symplectic.
- Letting z = (q, p), and defining J, the symplecticness condition is that the Jacobian matrix, B_s, of the mapping T_s satisfies

$$J = \begin{bmatrix} 0_{d\times d} & I_{d\times d} \\ -I_{d\times d} & 0_{d\times d} \end{bmatrix} \qquad B_s^T J^{-1} B_s = J^{-1}$$

- This implies volume conservation, since $\det(B_s^T)\det(J^{-1})\det(B_s) = \det(J^{-1})$ implies that $\det(B_s)^2$ is one

Discretizing Hamilton's Equations

- For computer implementation, Hamilton's equations must be approximated by discretizing time, using some small step size, ε.
- Starting with the state at time zero, we iteratively compute (approximately) the state at times ε, 2ε, 3ε, etc.

Discretizing Hamilton's Equations: Euler's Method

- Perhaps the best-known way to approximate the solution to a system of differential equations is Euler's method.

$$p_i(t+\varepsilon) = p_i(t) + \varepsilon \frac{dp_i}{dt}(t) = p_i(t) - \varepsilon \frac{\partial U}{\partial q_i}(q(t)), \tag{5.14}$$

$$q_i(t+\varepsilon) = q_i(t) + \varepsilon \frac{dq_i}{dt}(t) = q_i(t) + \varepsilon \frac{p_i(t)}{m_i}. \tag{5.15}$$

- Coming from:

$$\frac{dq_i}{dt} = [M^{-1}p]_i, \tag{5.6}$$

$$\frac{dp_i}{dt} = -\frac{\partial U}{\partial q_i}. \tag{5.7}$$

Discretizing Hamilton's Equations: Modification Euler's Method

- We simply use the *new* value for the momentum variables, *pi*, when computing the new value for the position variables, *qi*.

$$p_i(t+\varepsilon) = p_i(t) - \varepsilon \frac{\partial U}{\partial q_i}(q(t)), \tag{5.16}$$

$$q_i(t+\varepsilon) = q_i(t) + \varepsilon \frac{p_i(t+\varepsilon)}{m_i}. \tag{5.17}$$

Discretizing Hamilton's Equations: Leapfrog Method

- We start with a half step for the momentum variables
- then do a full step for the position variables, using the new values of the momentum variables
- finally do another half step for the momentum variables, using the new values for the position variables
- The leapfrog method preserves volume exactly,

$$p_i\,(t+\varepsilon/2) = p_i(t) - (\varepsilon/2)\,\frac{\partial U}{\partial q_i}(q(t)), \tag{5.18}$$

$$q_i(t+\varepsilon) = q_i(t) + \varepsilon\,\frac{p_i(t+\varepsilon/2)}{m_i}, \tag{5.19}$$

$$p_i(t+\varepsilon) = p_i\,(t+\varepsilon/2) - (\varepsilon/2)\,\frac{\partial U}{\partial q_i}(q(t+\varepsilon)). \tag{5.20}$$

Comparison:

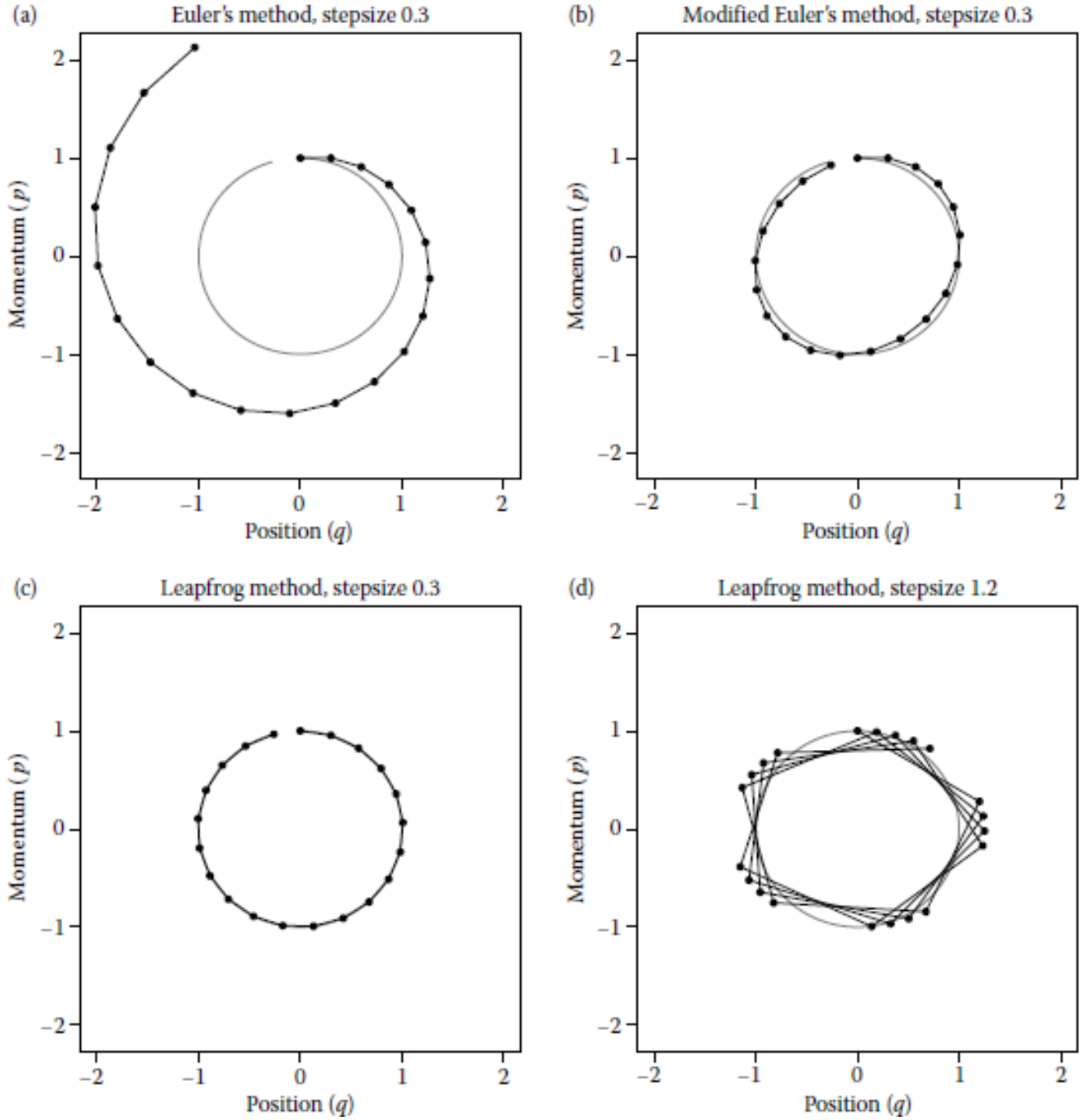
Results using three methods for approximating Hamiltonian dynamics, when:

$H(q, p) = q2/2 + p2/2.$

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).



Local and Global Error

- The ***local error*** is the error after one step, that moves from time t to time $t + \varepsilon$.
- The ***global error*** is the error after simulating for some fixed time interval, s, which will require s/ε steps.
- The Euler method and its modification above have order ε^2 local error and order ε global error.
- The leapfrog method has order ε^3 local error and order ε^2 global error.
- Claim: This difference is a consequence of leapfrog being reversible, since any reversible method must have global error that is of even order in ε.

MCMC from Hamiltonian Dynamics

- Using Hamiltonian dynamics to sample from a distribution requires translating the density function for this distribution to a potential energy function and introducing "momentum" variables to go with the original variables of interest (now seen as "position" variables).
- We can simulate a Markov chain
- each iteration resamples the momentum
- Metropolis update with a proposal found using Hamiltonian dynamics

Probability and the Hamiltonian: Canonical Distributions

- The distribution we wish to sample can be related to a potential energy function via the concept of a *canonical distribution* from statistical mechanics.
- Given some energy function, $E(x)$, for the state, x, of some physical system, the canonical distribution over states has probability or probability density function:

$$P(x) = \frac{1}{Z} \exp\left(\frac{-E(x)}{T}\right). \tag{5.21}$$

- T is the temperature of the system, and Z is the normalizing constant needed for this function to sum or integrate to one.
- For Hamiltonian:

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

- If $H(q, p) = U(q) + K(p)$, since q and p are independent:

$$P(q,p) = \frac{1}{Z} \exp\left(\frac{-U(q)}{T}\right) \exp\left(\frac{-K(p)}{T}\right), \tag{5.22}$$

The Hamiltonian Monte Carlo Algorithm

- We now have the background
- HMC can be used to sample only from continuous distributions on R^d for which the density function can be evaluated (perhaps up to an unknown normalizing constant).
- Assume that the density is nonzero everywhere.
- We must also be able to compute the partial derivatives of the log of the density function.
- These derivatives must therefore exist, except perhaps on a set of points with probability zero, for which some arbitrary value could be returned.

The Hamiltonian Monte Carlo Algorithm

- HMC samples from the canonical distribution for q and p defined by

$$P(q,p) = \frac{1}{Z} \exp\left(\frac{-U(q)}{T}\right) \exp\left(\frac{-K(p)}{T}\right), \qquad (5.22)$$

- q has the distribution of interest, as specified using the potential energy function $U(q)$.We can choose the distribution of the momentum variables, p, which are independent of q, as we wish, specifying the distribution via the kinetic energy function, $K(p)$.
- Current practice with HMC is to use a quadratic kinetic energy:

$$K(p) = p^T M^{-1} p/2. \qquad (5.5)$$

- which leads p to have a zero-mean multivariate Gaussian distribution.
- The kinetic energy function producing this distribution (setting $T = 1$) is:

$$K(p) = \sum_{i=1}^{d} \frac{p_i^2}{2m_i}. \qquad (5.23)$$

The Two Steps of the HMC Algorithm

- The first changes only the momentum
  - new values for the momentum variables are randomly drawn from their Gaussian distribution
  - independently of the current values of the position variables
- The second may change both position and momentum
  - In the second step, a Metropolis update is performed, using Hamiltonian dynamics (e.g., leapfrog method) to propose a new state.

$$p_i(t + \varepsilon/2) = p_i(t) - (\varepsilon/2)\,\frac{\partial U}{\partial q_i}(q(t)), \tag{5.18}$$

$$q_i(t + \varepsilon) = q_i(t) + \varepsilon\,\frac{p_i(t + \varepsilon/2)}{m_i}, \tag{5.19}$$

$$p_i(t + \varepsilon) = p_i(t + \varepsilon/2) - (\varepsilon/2)\,\frac{\partial U}{\partial q_i}(q(t + \varepsilon)). \tag{5.20}$$

- Both steps leave the canonical joint distribution of *(q, p)* invariant, and hence their combination also leaves this distribution invariant.

The Two Steps of the HMC Algorithm (Cont.)

- Starting with the current state, *(q, p)*, Hamiltonian dynamics is simulated for *L* steps using the leapfrog method (or some other reversible method that preserves volume), with a step-size of ε.
- Here, *L* and ε are parameters of the algorithm, which need to be tuned to obtain good performance.
- The momentum variables at the end of this *L*-step trajectory are then negated, giving a proposed state (q^*, p^*).
- This proposed state is accepted as the next state of the Markov chain with probability:

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

- If the proposed state is not accepted (i.e., it is rejected), the next state is the same as the current state

Ergodicity of HMC

- The HMC algorithm will also be “ergodic”
- It will not be trapped in some subset of the state
- Hence will asymptotically converge to its (unique) invariant distribution

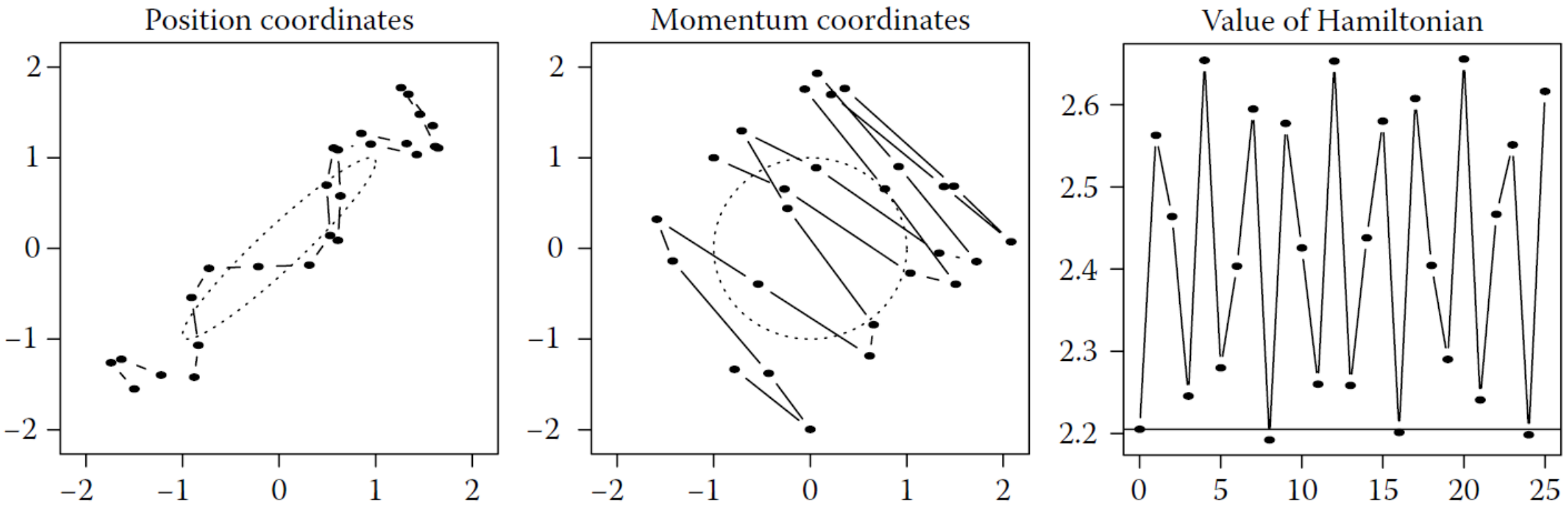
Illustrations of HMC and Its Benefits: Trajectories for a Two-Dimensional Problem

- Position variables: Consider sampling from a distribution for two variables that is bivariate Gaussian.
- With means of zero, standard deviations of one, and correlation 0.95.
- Momentum variables: defined to have a Gaussian distribution
- With means of zero, standard deviations of one, and zero correlation

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

- See the results in next slide

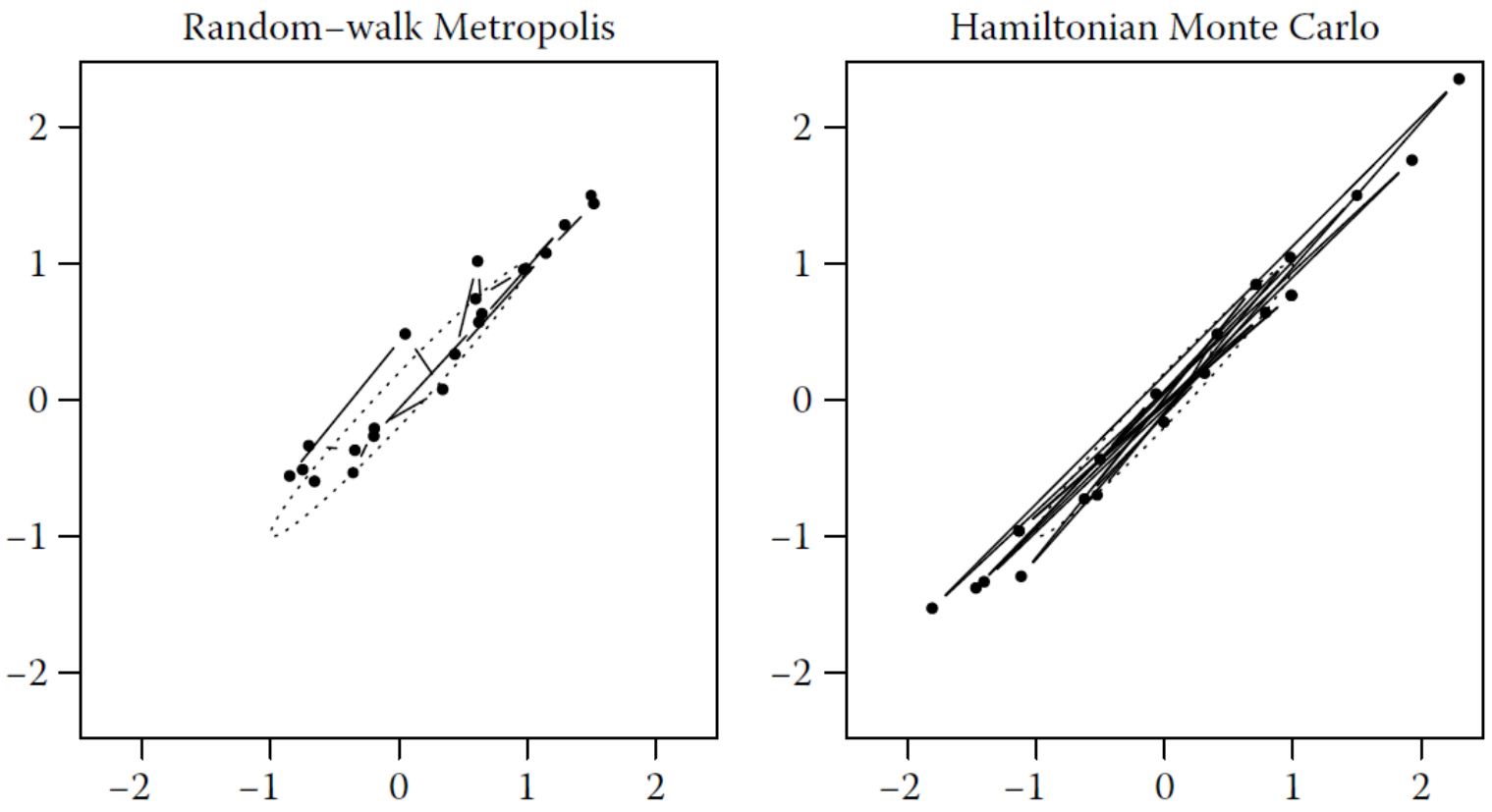
A trajectory for a two-dimensional Gaussian distribution, simulated using 25 leapfrog steps with a step-size of 0.25. The ellipses plotted are one standard deviation from the means. The initial state had $q = [-1.50, -1.55]^T$ and $p = [-1, 1]^T$.



Notice that this trajectory does not resemble a random walk. Instead, starting from the lower left-hand corner, the position variables systematically move upward and to the right, until they reach the upper right-hand corner, at which point the direction of motion is reversed. The consistency of this motion results from the role of the momentum variables.

Compare with random walk

- Same example but with stronger correlation of 0.98.
- Compare with random-walk
- The HMC rejection rate for these trajectories was 0.09.
- The rejection rate for these random-walk proposals was 0.37.



The Benefit of Avoiding Random-Walks

- Avoidance of random-walk behavior, as illustrated above, is one major benefit of HMC.

- Claim: Because the random-walk Metropolis proposals have no tendency to move consistently in the same direction, we would need around 100 iterations of random-walk Metropolis in which the proposal was accepted to move to a nearly independent state.

Tuning HMC

One practical impediment to the use of Hamiltonian Monte Carlo is the need to select suitable values for

- The leapfrog step-size, ε,
  - Too large a step-size will result in a very low acceptance rate for states proposed by simulating trajectories.
  - Too small a step-size will either waste computation time
- The number of leapfrog steps, L, which together determine the length of the trajectory in fictitious time, εL.
  - Setting the trajectory length by trial and error therefore seems necessary.
- Claim: Tuning HMC is more difficult in some respects than tuning a simple Metropolis method.

Scaling with Dimensionality

- For problems in which the dimensionality is moderate to high, another benefit of HMC over simple random-walk Metropolis methods is a slower increase in the computation time needed as the dimensionality increases.

Thank you