An Introduction to Hamiltonian Monte Carlo Methods

Siddharth Sabata
University of California Santa Barbara

Markov Chain Monte Carlo Methods

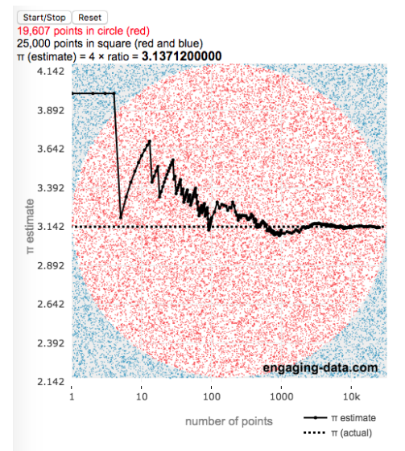
Markov Chain Monte Carlo methods are a type of algorithm that allow us to sample from a probability distribution without knowing what the distribution looks like. We'll begin by defining the individual parts that make up MCMC: Markov Chains and Monte Carlo methods.

Markov Chains A Markov Chain is a random process that undergoes state changes. The chain also has a property, called the Markov property, where the probability of moving to the next state is only dependent on the current state. Another way to say this is "in order to know the future, the knowledge of the past does not add anything to the knowledge of the present" ([1, p. 45]).

More formally, the discrete, or finite, case for a Markov Chain is defined as a discrete-time process $\{X_n\}_{n\geq 0}$, i.e. a collection of random variables with the index n usually representing time, with values in a countable space E is a Markov chain if for all $n \geq 0$ and all states $i_0, i_1, \ldots, i_{n-1}, i, j \in E$,

$$\mathbb{P}(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = \mathbb{P}(X_{n+1} = j | X_n = i),$$

where $\mathbb{P}(Y = j | X = i)$ means the conditional probability of the event $\{Y = j\}$ given $\{X = i\}$ ([1, p. 46 (Definition 2)]).



Monte Carlo Methods Monte Carlo methods are algorithms that estimate quantities that are too difficult to obtain analytically or by discretization. This is done through repeated random sampling and averaging the results. Some common examples are estimating π or integrals.

Markov Chain Monte Carlo Algorithm MCMC methods were proposed by Metropolis and his colleagues in 1953. They wanted to develop a method to estimate an unknown target distribution using Markov Chains.

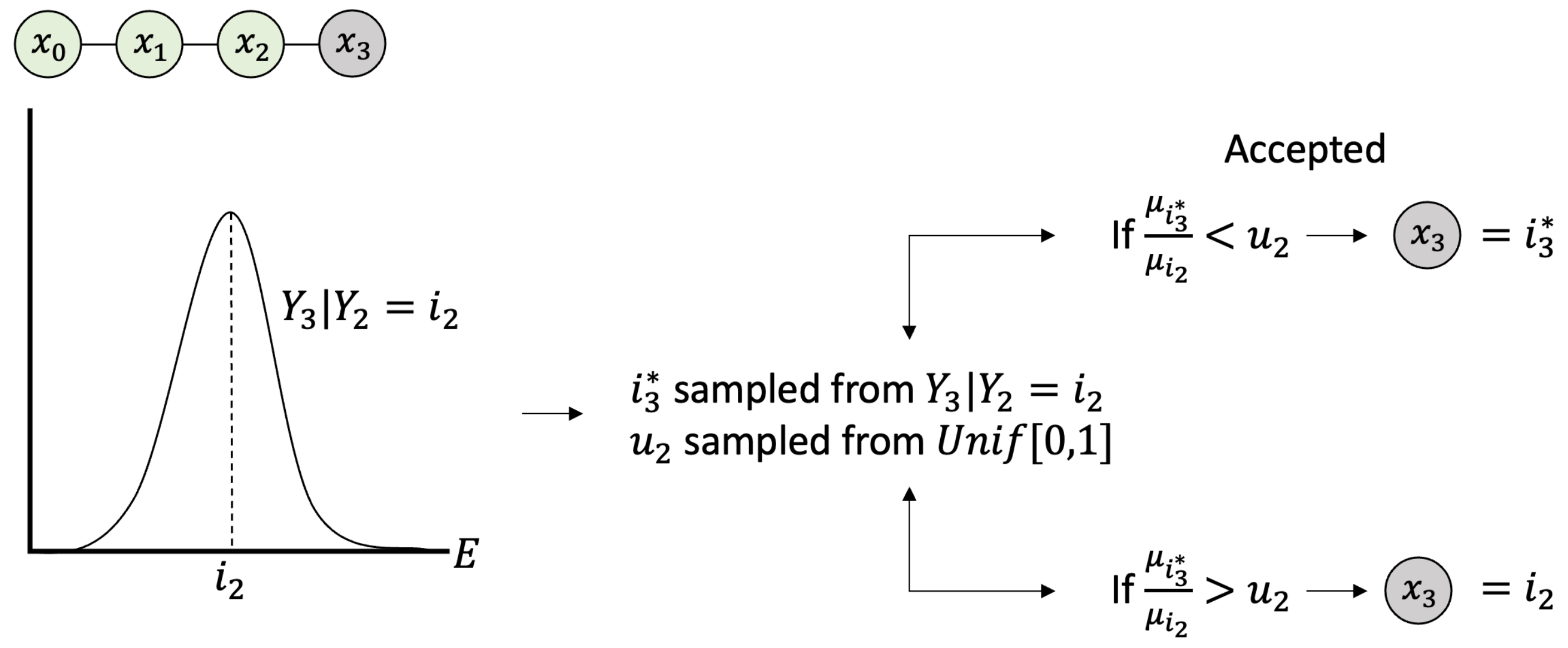
The algorithm for the discrete, or finite, case is defined as the following:

- Choose a value for i_0 for X_0 (randomly or e.g $i_0 = 0$).
- Once values $i_0, \ldots, i_n$ of $X_0, \ldots, X_n$, respectively, have been found:
  - Generate a proposed value $i^*_{n+1} \in E$ from an auxiliary distribution $Y_{n+1} | Y_n = i_n$.
  - If $\mu_{i^*_{n+1}} / \mu_{i_n} > u_n$ set $X_{n+1} = i^*_{n+1}$; in this case we say that the proposal is accepted. Else set $X_{n+1} = i_n$ and we say the proposal is rejected.

Here, E is a discrete state space, i.e., the set of all possible events, $\mu_{i^*_{n+1}}$ and μ_{i_n} are the value of the target distribution at i^*_{n+1} and i_n respectively, and u_n is drawn from the uniform distribution on $[0, 1]$ to carry out acceptance/rejection step with probability $\min\{1, \ \mu_{i^*_{n+1}} / \mu_{i_n}\}$ ([1, p. 55]).

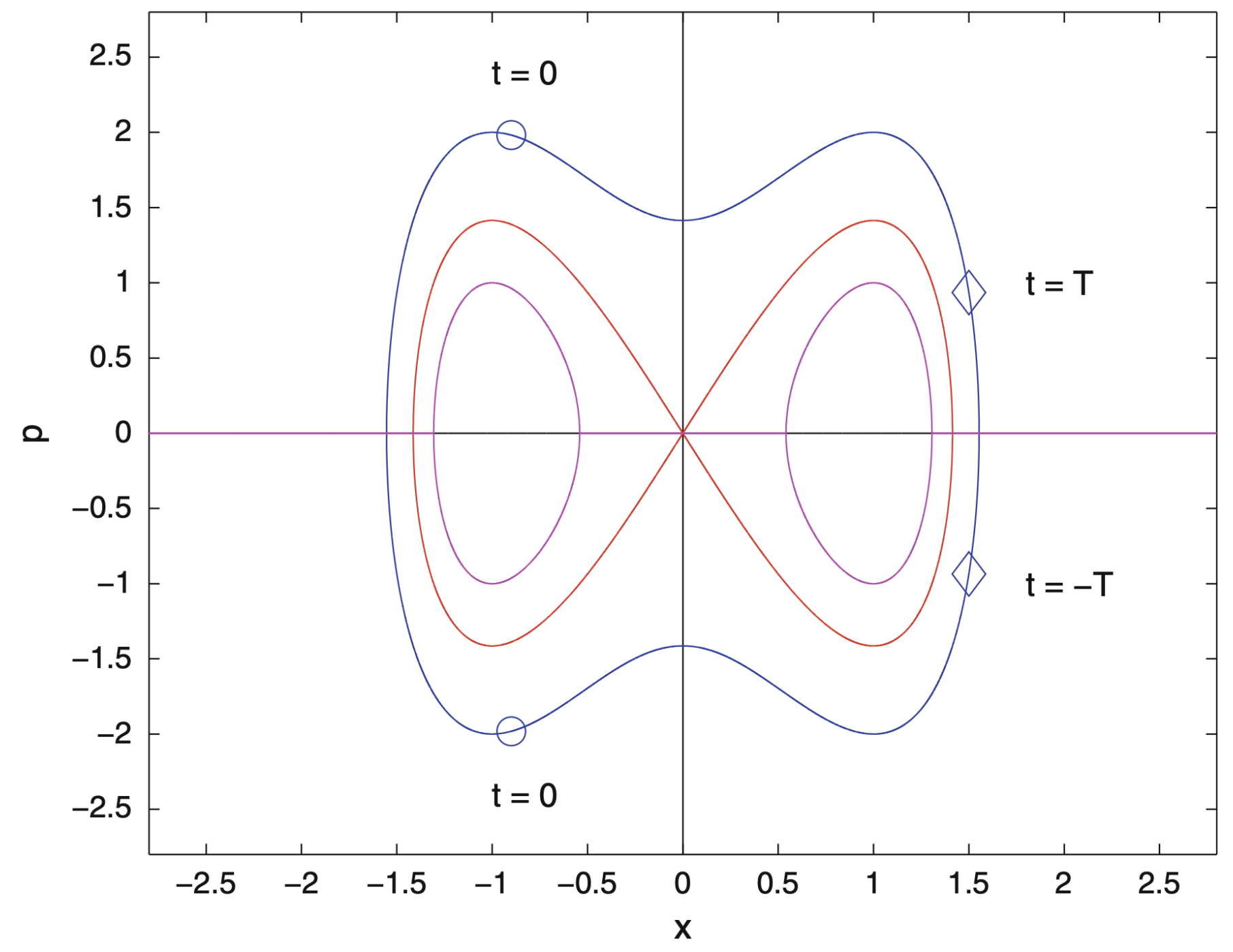
Typically, first few hundreds or thousands of samples will not be close to those drawn from the target distribution. Once the Markov Chain converges to what is called its *stationary distribution*, a point at which the probability distribution representing the movement from one state to the next does not change, the Markov Chain transition is just like "independent sampling" from our target distribution μ.

Visual Representation of MCMC Algorithm


x_0
x_1
x_2
x_3
$Y_3|Y_2 = i_2$
i_2
E
i^*_3 sampled from $Y_3|Y_2 = i_2$
u_2 sampled from $Unif[0,1]$
Accepted
If $\frac{\mu_{i^*_3}}{\mu_{i_2}} < u_2$
x_3
$= i^*_3$
If $\frac{\mu_{i^*_3}}{\mu_{i_2}} > u_2$
x_3
$= i_2$

Hamiltonian Dynamics

Visual Representation of the Phase Space [1, p. 74 (Figure 13)]



Hamiltonian Dynamics Hamiltonian dynamics is another way to look at classical mechanics in physics. This framework describes how a system changes over time based on the Hamiltonian function H which represents the total mechanical energy of a system, which equals the sum of the kinetic and potential energy (represented by V) of the system of ν particles:

$$H = \sum_i \frac{1}{2m_i} \mathbf{p}_i^T \mathbf{p}_i + V(\mathbf{r}_1, \ldots, \mathbf{r}_\nu),$$

where $\mathbf{r_i}$, $\mathbf{p_i}$ and m_i correspond to the position, momentum, and mass of i-th particle ([1, p. 70]).

Hamiltonian Equations From now on, we discuss the case with only one particle for simplicity of presentation. In the phase space space $\mathbb{R}^D$, $D = 2d$, $(\mathbf{p}, \mathbf{x}) \in \mathbb{R}^D$, to each smooth real-valued function $H = H(\mathbf{p}, \mathbf{x})$ (Hamiltonian), the corresponding system of first order differential equations, called the *canonical* or *Hamilton's* equations describes the time evolution of the system ([1, p. 71]):

$$\frac{d}{dt} p_j = -\frac{\partial H}{\partial x_j}, \quad \frac{d}{dt} x_j = +\frac{\partial H}{\partial p_j}, \quad j = 1, \ldots, d.$$

Flow The flow of a Hamiltonian system is denoted with $\{\Phi_t\}_{t \in \mathbb{R}}$. Φ_t is a map in the phase space, $\Phi_t : \mathbb{R}^D \to \mathbb{R}^D$, that is defined as follows: $\Phi_t(\mathbf{p}_0, \mathbf{x}_0)$ is the solution $(\mathbf{p}(t), \mathbf{x}(t))$ at time t of the canonical equation with the initial value $(\mathbf{p}_0, \mathbf{x}_0)$ at $t = 0$. Basically, Φ_t tells us how the system evolves over time ([1, p. 71-72]).

Property 1: Conservation of Energy The function H is a conserved quantity of the Hamilton's equations. Along solutions, we have

$$\frac{d}{dt} H(\mathbf{p}(t), \mathbf{x}(t)) = \sum_j \left(\frac{\partial H}{\partial p_j} \frac{d}{dt} p_j + \frac{\partial H}{\partial x_j} \frac{d}{dt} x_j \right) = \sum_j \left(-\frac{\partial H}{\partial p_j} \frac{\partial H}{\partial x_j} + \frac{\partial H}{\partial x_j} \frac{\partial H}{\partial p_j} \right) = 0.$$

Therefore,

$$H(\mathbf{p}(t), \mathbf{x}(t)) = H(\mathbf{p}(0), \mathbf{x}(0)).$$

In other words, when $d = 1$, the dynamics follows a contour curve of H on the phase space [1, p. 72] (see the figure above).

Property 2: Conservation of Volume If we take the area of a specific region in the phase space and evolve it through time, the area remains the same even though the shape of the area might change ([1, p. 72-73]).

Property 3: Reversibility This property states that if we go backwards in the time evolution of our system, the system's motion will also go backwards. This means that if we knew the state of the system at any point in time, we could run it backwards and find the state of the system at a previous point in time ([1, p. 73]).

These properties lead to many desirable aspects, when combined with the MCMC, including the existence of a stationary distribution and high acceptance rate of proposed states.

Hamiltonian Monte Carlo

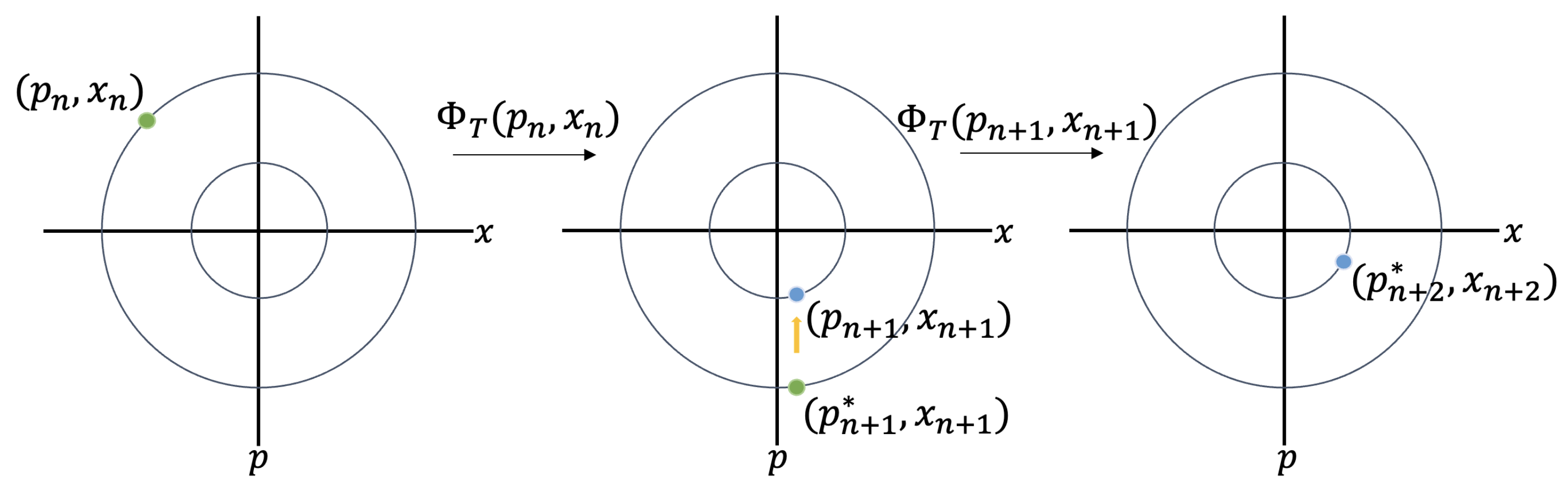
We can combine MCMC and Hamiltonian dynamics together to form Hamiltonian Monte Carlo methods. These methods use Hamiltonian dynamics for the Markov chain transition. We write the target density $\pi(\mathbf{x})$ in the state space $\mathbb{R}^D$ as $\exp(-V(\mathbf{x}))$. We can think of $\mathbf{x} \in \mathbb{R}^d$ as the position of our mechanical system, $V(\mathbf{x})$ as corresponding potential energy, and $\mathbf{p} \in \mathbb{R}^d$ as momentum. $\mathbf{p}$ is usually endowed a Gaussian density with the identity covariance matrix, which corresponds to mass being 1 in the canonical equation. With this set up, we can construct a Markov Chain in $\mathbb{R}^d$ ([1, p. 78]). For the simplicity of presentation, we further assume $d = 1$ below.

HMC Algorithm (analytic flow version) Define the transitions $x_n \mapsto x_{n+1}$ in the state space $\mathbb{R}^d$ by the following procedure:

- Draw p_n from a Gaussian density.
- Find $(p^*_{n+1}, x_{n+1}) = \Phi_T(p_n, x_n)$, where Φ_T is the T-flow of the canonical system (the Hamiltonian equations) with Hamiltonian function H.

Then $x_n \mapsto x_{n+1}$ defines a Markov chain in $\mathbb{R}^d$ that has the target $\pi(x) \propto$ exp$(-V(x))$ as an invariant probability distribution ([1, p. 78-79 (Theorem 8)]).

Visual Representation of Sampling from the Hamiltonian Phase Space



p_n sampled from Gaussian density
x_n from current state

p^*_{n+1} discarded
p_{n+1} sampled from Gaussian density
x_{n+1} is new state

p^*_{n+2} discarded
x_{n+2} is new state

Why is this helpful? MCMC algorithms are typically inefficient due to low acceptance rates, resulting in the Markov Chain converging to its stationary distribution taking a long time. On top of this, MCMC algorithms usually suffer from high autocorrelation which comes from each sample being too close to the previous one. This also leads to the Markov Chain taking a longer time to converge to its stationary distribution.

Notice that the momentum p^*_{n+1} is refreshed every iteration. This makes it possible to explore different energy levels, which likely leads to wider range of state space. Also note that HMC does not even feature acceptance/rejection step because the acceptance rate is always 1: the ratio of the current and proposed energy levels. This dramatically reduces autocorrelation, which is high when the states are repeated as is the case when the proposed states are often rejected.

References

[1] J. M. Sanz-Serna. "Markov chain Monte Carlo and numerical differential equations". In: *Current challenges in stability issues for numerical differential equations.* Vol. 2082. Lecture Notes in Math. Springer, Cham, 2014, pp. 39–88.

Acknowledgements

Thank you to Jea-Hyun Park for his mentorship and guidance and the UC Santa Barbara Directed Reading Program for this opportunity.