

# STAT 238 - Bayesian Statistics

## Lecture Thirty Four

Spring 2026, UC Berkeley

Aditya Guntuboyina

22 April 2026

### 1 Hamiltonian Monte Carlo

There is a given target probability density  $\pi$  and our goal is to construct a Markov chain which satisfies detailed balance with respect to  $\pi$ . Given a current value  $x$ , HMC constructs the next value  $y$  by solving the following second order ODE:

$$\ddot{x}(t) = \nabla \log \pi(x(t)) \quad \text{with initialization } x(0) = x \text{ and } \dot{x}(0) = z \sim N(0, I_d). \quad (1)$$

The ODE (1) can be alternatively written in the following first order form:

$$\begin{pmatrix} \dot{x}(t) \\ \dot{v}(t) \end{pmatrix} = \begin{pmatrix} v(t) \\ \nabla \log \pi(x(t)) \end{pmatrix} \quad \text{with initialization } \begin{pmatrix} x(0) \\ v(0) \end{pmatrix} = \begin{pmatrix} x \\ z \end{pmatrix} \quad (2)$$

which can also be written in terms of the Hamiltonian:

$$H(x, v) = H(x_1, \dots, x_d, v_1, \dots, v_d) := -\log \pi(x) + \frac{1}{2} \|v\|^2 \quad (3)$$

as

$$\begin{pmatrix} \dot{x}(t) \\ \dot{v}(t) \end{pmatrix} = \begin{pmatrix} \frac{\partial H}{\partial v}(x(t), v(t)) \\ -\frac{\partial H}{\partial x}(x(t), v(t)) \end{pmatrix} \quad \text{with initialization } \begin{pmatrix} x(0) \\ v(0) \end{pmatrix} = \begin{pmatrix} x \\ z \end{pmatrix} \quad (4)$$

Most authors (see e.g., Neal [1]) use the formulation (4) to describe HMC. The advantage of the formulation (1) is that it makes the connection to MALA very clear (essentially, MALA proposals are obtained by replacing  $\nabla \log \pi(x(t))$  by  $\nabla \log \pi(x(0))$ ).

On the other hand, the main advantage of the Hamiltonian formulation is that it also works for choices of the Hamiltonian that are different from (3). For example, consider the alternative choices of the Hamiltonian:

1.  $H(x, v) = -\log \pi(x) + v^T M v / 2$  for a positive definite matrix  $M$ . Here the dynamics (4) become

$$\begin{pmatrix} \dot{x}(t) \\ \dot{v}(t) \end{pmatrix} = \begin{pmatrix} Mv(t) \\ -\nabla \log \pi(x(t)) \end{pmatrix}.$$

because  $\frac{\partial H}{\partial x} = -\nabla \log \pi(x)$  and  $\frac{\partial H}{\partial v} = Mv$ . This formulation of the Hamiltonian is meaningful when the variables  $x_1, \dots, x_d$  have different scales. One can also make  $M$  depend on  $x$ . This is related to Riemannian Manifold HMC (see Girolami and Calderhead [2]).

2.  $H(x, v) = -\log \pi(x) + \|v\|_1$ . Here the dynamics (4) become

$$\begin{pmatrix} \dot{x}(t) \\ \dot{v}(t) \end{pmatrix} = \begin{pmatrix} \text{sign}(v(t)) \\ -\log \pi(x(t)) \end{pmatrix}$$

because  $\frac{\partial H}{\partial v} = \text{sign}(v)$ . Also note that  $v(t) = (v_1(t), \dots, v_d(t))$  and  $\text{sign}(v(t))$  is interpreted coordinatewise as  $(\text{sign}(v_1(t)), \dots, \text{sign}(v_d(t)))$ . These are interesting dynamics that cannot be written in a second order form as in (1).

## 2 Hamiltonian Dynamics

Some understanding of Hamiltonian Dynamics will be useful for HMC. Hamiltonian Dynamics refers to the ODE (4)

$$\begin{pmatrix} \dot{x}(t) \\ \dot{v}(t) \end{pmatrix} = \begin{pmatrix} \frac{\partial H}{\partial v}(x(t), v(t)) \\ -\frac{\partial H}{\partial x}(x(t), v(t)) \end{pmatrix} \text{ with initialization } \begin{pmatrix} x(0) \\ v(0) \end{pmatrix} = \begin{pmatrix} x \\ v \end{pmatrix} \quad (5)$$

The above is the same as (4) except that we denote the initial velocity by  $v$  (instead of  $v$  in (4)). In this section, we will not use the specific form of the Hamiltonian given by (3), but we will take the more general form:

$$H(x, v) = -\log \pi(x) + K(v) \quad (6)$$

where  $K(v)$  is a symmetric function of  $v$  i.e.,  $K(-v) = K(v)$ .

We will view the Hamiltonian dynamics (run for a fixed time  $\sigma$ ) as a mapping from  $(x, v)$  (the initial state) to a final state  $(x(\sigma), v(\sigma))$ . We will denote this mapping by  $T_\sigma(x, v)$ :

$$T_\sigma(x, v) = (x(\sigma), v(\sigma)).$$

The space of all pairs of position and velocity  $(x, v)$  will be referred to as the phase space.

The basic properties of  $T_\sigma(\cdot)$  that we need for HMC are summarized in this section. For more details, you can refer to Neal [1].

### 2.1 Property One: Invertibility or Reversibility

The map  $T_\sigma(x, v)$  is invertible and its inverse can be written in a simple manner. Let  $S(x, v) = (x, -v)$  be the operator which flips the velocity. Then

$$T_\sigma^{-1} = ST_\sigma S. \quad (7)$$

In other words, to compute  $T_\sigma^{-1}$  at a point  $(y, w)$ , we need to take the following three steps:

1. First, flip the velocity to obtain  $(y, -w)$ .
2. Second, run the Hamiltonian dynamics for time  $\sigma$  starting at  $(y, -w)$  to obtain  $(x, -v)$  at time  $\sigma$ .
3. Third, flip velocity again to obtain  $(x, v)$ .

The formula (7) has the following simple consequence:

$$T_\sigma S = (T_\sigma S)^{-1}. \quad (8)$$

To see this, simply note

$$(T_\sigma S)^{-1} = S^{-1}T_\sigma^{-1} = S^{-1}ST_\sigma S = T_\sigma S.$$

A function  $g$  for which  $g^{-1} = g$  (which is equivalent to  $g(g(x)) = x$ ) is called an *involution*. The above statement is therefore equivalent to saying that  $T_\sigma S$  is an involution. Involutions have been recently used to unify many MCMC algorithms (see e.g., Glatt-Holtz et al. [4]).

## 2.2 Property Two: Hamiltonian Conservation

Hamiltonian dynamics conserve the Hamiltonian in the sense that:

$$H(x(t), v(t)) = \text{constant}. \quad (9)$$

To see this, just note that

$$\begin{aligned} \frac{d}{dt}H(x(t), v(t)) &= \sum_{i=1}^d \left( \frac{\partial H}{\partial x_i} \dot{x}_i(t) + \frac{\partial H}{\partial v_i} \dot{v}_i(t) \right) \\ &= \sum_{i=1}^d \left( \frac{\partial H}{\partial x_i} \frac{\partial H}{\partial v_i} + \frac{\partial H}{\partial v_i} \left( -\frac{\partial H}{\partial x_i} \right) \right) = \sum_{i=1}^d \left( \frac{\partial H}{\partial x_i} \frac{\partial H}{\partial v_i} - \frac{\partial H}{\partial v_i} \frac{\partial H}{\partial x_i} \right) = 0 \end{aligned}$$

## 2.3 Property Three: Volume Preservation

Volume preservation means that if we take a region  $R$  in the phase space  $(x, v)$ , then

$$\text{vol}(R) = \text{vol}(T_\sigma(R)) \quad (10)$$

where  $T_\sigma(R) = \{T_\sigma(x, v) : (x, v) \in R\}$ .

The volume preservation property (10) is equivalent to

$$\text{vol}(R) = \text{vol}(T_\sigma(R)) = \int I\{(y, w) \in T_\sigma(R)\}d(y, w) = \int I\{T_\sigma^{-1}(y, w) \in R\}d(y, w).$$

Using the change of variable  $(x, v) = T_\sigma^{-1}(y, w)$  above, we get

$$\int I\{T_\sigma^{-1}(y, w) \in R\}d(y, w) = \int I\{(x, v) \in R\}|\det JT_\sigma(x, v)|d(x, v)$$

where  $JT_\sigma$  denotes the Jacobian of  $T_\sigma$ . Thus volume preservation can be proved by showing that the Jacobian  $JT_\sigma$  has determinant equal to 1 for all  $(x, v)$ :

$$\det JT_\sigma(x, v) = 1 \quad \text{for all } (x, v). \quad (11)$$

Here is a sketch of the proof of (11). Assume first that  $\sigma$  is small so that  $T_\sigma$  can be well-approximated by a linear function:

$$T_\sigma(x, v) \approx \begin{pmatrix} x \\ v \end{pmatrix} + \sigma \begin{pmatrix} \frac{\partial H}{\partial v}(x, v) \\ -\frac{\partial H}{\partial x}(x, v) \end{pmatrix}$$

As a result

$$JT_\sigma \approx I + \sigma \begin{pmatrix} \frac{\partial^2 H}{\partial x \partial v} & \frac{\partial^2 H}{\partial v^2} \\ -\frac{\partial^2 H}{\partial x^2} & -\frac{\partial^2 H}{\partial x \partial v} \end{pmatrix}$$

which shows that

$$\det(JT_\sigma) \approx \det \begin{pmatrix} 1 + \sigma \frac{\partial^2 H}{\partial x \partial v} & \sigma \frac{\partial^2 H}{\partial v^2} \\ -\sigma \frac{\partial^2 H}{\partial x^2} & 1 - \sigma \frac{\partial^2 H}{\partial x \partial v} \end{pmatrix} = 1 + O(\sigma^2).$$

If  $\sigma$  is not small, then break  $[0, \sigma]$  into  $N$  intervals each of length  $\sigma/N$ , and then decompose  $T_\sigma$  as  $T_N \circ T_{N-1} \circ \dots \circ T_1$  where  $T_j$  is the mapping corresponding to Hamiltonian dynamics from  $t = (j-1)\sigma/N$  to  $t = j\sigma/N$ . For each  $j$ , we use the previous argument to claim  $\det JT_j = 1 + O(\sigma^2/N^2)$ . Taking the product over  $j$ , we get

$$\det JT_\sigma = \prod_{j=1}^N (1 + O(\sigma^2/N^2)) \rightarrow 1$$

as  $N \rightarrow \infty$ . This completes the heuristic argument for (11).

### 3 Discretization of (2)

The ODE (2) cannot be solved exactly for most  $\pi(\cdot)$ . So we have to approximately solve it using a discretization technique. We fix a step size  $\epsilon$  and approximate the ODE (2) at  $t = 0, \epsilon, 2\epsilon, \dots$ . More specifically, we will construct:

$$\begin{pmatrix} x(0) \\ v(0) \end{pmatrix}, \begin{pmatrix} x(\epsilon) \\ v(\epsilon) \end{pmatrix}, \begin{pmatrix} x(2\epsilon) \\ v(2\epsilon) \end{pmatrix}, \dots$$

Below we will describe how to obtain  $(x(t+\epsilon), v(t+\epsilon))$  from  $(x(t), v(t))$  (this process will then be iteratively applied starting at  $t = 0$ ).

The standard approach to discretization of the HMC equation (2) is to use the leapfrog discretization. The leapfrog discretization uses the following formulae to construct  $x(t+\epsilon)$  and  $v(t+\epsilon)$  from  $x(t), y(t)$ .

$$\begin{aligned} v\left(t + \frac{\epsilon}{2}\right) &= v(t) + \frac{\epsilon}{2} \nabla \log \pi(x(t)) \\ x(t + \epsilon) &= x(t) + \epsilon v\left(t + \frac{\epsilon}{2}\right) \\ v(t + \epsilon) &= v\left(t + \frac{\epsilon}{2}\right) + \frac{\epsilon}{2} \nabla \log \pi(x(t + \epsilon)) \end{aligned} \tag{12}$$

If we denote this map from  $(x(t), y(t))$  to  $(x(t+\epsilon), y(t+\epsilon))$  by  $T_\epsilon^{\text{disc}}(x, v)$ , then it is straightforward to verify that

$$\begin{aligned} T_\epsilon^{\text{disc}}(x, v) &= \left( x + \epsilon v + \frac{\epsilon^2}{2} \nabla \log \pi(x), v + \frac{\epsilon}{2} \nabla \log \pi(x) + \frac{\epsilon}{2} \nabla \log \pi\left(x + \epsilon v + \frac{\epsilon^2}{2} \nabla \log \pi(x)\right) \right). \end{aligned} \tag{13}$$

Note that the position update above is reminiscent of MALA.

If we apply  $N$  leapfrog steps in succession, then the overall mapping is given by:

$$T_\epsilon^{\text{disc}, N} = T_\epsilon^{\text{disc}} \circ \dots \circ T_\epsilon^{\text{disc}}.$$

This function  $T_\epsilon^{\text{disc}, N}$  shares the following two properties of the continuous Hamiltonian dynamics:

1. **Invertibility or Time Reversibility:** One can check that  $ST_\epsilon^{\text{disc},N}S$  serves as the inverse of  $T_\epsilon^{\text{disc},N}$ . This can be verified by proving that

$$\left(T_\epsilon^{\text{disc}}\right)^{-1} = ST_\epsilon^{\text{disc}}S.$$

Given the explicit formula for  $T_\epsilon^{\text{disc}}$  in (13), this verification is straightforward.

2. **Volume Preserving:**  $T_\epsilon^{\text{disc},N}$  is volume-preserving. This can be verified by showing the the determinant of the Jacobian of  $T_\epsilon^{\text{disc},N}$  equals one. This, in turn, can be verified by proving that the determinant of the Jacobian of  $T_\epsilon^{\text{disc}}$  equals 1. This can be verified directly by using the formula (13) or by noting (from (12)) that  $T_\epsilon^{\text{disc}}$  is the composition of three mappings:

$$T_\epsilon^{\text{disc}} = A_{\epsilon/2} \circ B_\epsilon \circ A_{\epsilon/2}$$

where

$$A_{\epsilon/2}(x, v) = \left(x, v + \frac{\epsilon}{2} \nabla \log \pi(x)\right) \quad \text{and} \quad B_\epsilon(x, v) = (x + \epsilon v, v).$$

These mappings are very simple and one can directly verify that they are volume preserving (i.e., the determinant of their Jacobians equals 1).

While the discretized Hamiltonian dynamics satisfy invertibility (or time-reversibility) and volume preservation, they do not satisfy Hamiltonian conservation (unlike continuous Hamiltonian dynamics). Because of this, a Metropolis acceptance correction has to be applied when implementing HMC with leapfrog discretization. This will be described in the next lecture.

## 4 On stationarity of $\pi$ for the continuous Hamiltonian Dynamics

For the continuous Hamiltonian dynamics given by (4), it turns that if the initial conditions  $x$  and  $v$  are distributed independently according to  $x \sim \pi$  and  $v \sim N(0, I_d)$ , then  $x(\sigma)$  and  $v(\sigma)$  are also independently distributed as  $x(\sigma) \sim \pi$  and  $v(\sigma) \sim N(0, I_d)$  for every  $\sigma$ . This can be proved using the continuity equation that we discussed in the last lecture. This argument is given below.

### 4.1 Continuity equation

Consider the ODE

$$\dot{X}(t) = V(t, X(t)) \quad \text{for } t \geq 0. \tag{14}$$

Here  $X(t) \in \mathbb{R}^d$  and  $V(t, \cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^d$ . Suppose we initialize the ODE at  $t = 0$  with a random variable having density  $\rho_b$ :

$$X(0) \sim \rho_b.$$

What then is the density of  $X(t)$ ? Let  $\rho(t, x)$  denote the density of  $X(t)$  evaluated at  $x$ . Then  $\rho(t, x)$  satisfies the following PDE:

$$\frac{\partial}{\partial t} \rho(t, x) = -\nabla \cdot (V(t, x) \rho(t, x)) \quad \text{with initialization } \rho(0, x) = \rho_b(x), \tag{15}$$

where

$$\nabla \cdot (V(t, x)\rho(t, x)) = \operatorname{div}(V(t, x)\rho(t, x)) := \sum_{i=1}^d \frac{\partial}{\partial x_i} (V_i(t, x)\rho(t, x))$$

The PDE (15) is known as the continuity equation or Transport equation, Fokker-Planck equation (it is closely related to the Fokker-Planck and Kolmogorov Forward equations).

The continuity equation has been popular in the recent literature on generative modeling (see e.g., Lai et al. [3, Equation (5.2.8), Theorem 5.2.2, Section B.1.2]).

## 4.2 Application to Hamiltonian Monte Carlo

Observe that (4) is a special case of (14) with  $x$  in (14) corresponding to  $(x, v)$  and

$$V(t, x, v) = \begin{pmatrix} \frac{\partial H}{\partial v} \\ -\frac{\partial H}{\partial x} \end{pmatrix}$$

Suppose now that  $x \sim \pi$  so that the initial density of  $(x(0), v(0))$  is

$$\rho_b(x, v) = \pi(x)\phi(v) = (2\pi)^{-d/2} \exp(-H(x, v)).$$

We then claim that

$$\rho(t, x, v) = (2\pi)^{-d/2} \exp(-H(x, v)) \quad \text{for every } t \geq 0$$

satisfies the continuity equation (15). To see this, simply note that

$$\begin{aligned} \nabla \cdot (V\rho) &= (2\pi)^{-d/2} \nabla \cdot \left( e^{-H} \begin{pmatrix} \frac{\partial H}{\partial v} \\ -\frac{\partial H}{\partial x} \end{pmatrix} \right) \\ &= (2\pi)^{-d/2} \sum_{i=1}^d \left[ \frac{\partial}{\partial x_i} \left( e^{-H} \frac{\partial H}{\partial v_i} \right) - \frac{\partial}{\partial v_i} \left( e^{-H} \frac{\partial H}{\partial x_i} \right) \right] \\ &= (2\pi)^{-d/2} \sum_{i=1}^d \left[ e^{-H} \left( \frac{\partial^2 H}{\partial x_i \partial v_i} - \frac{\partial H}{\partial x_i} \frac{\partial H}{\partial v_i} \right) - e^{-H} \left( \frac{\partial^2 H}{\partial x_i \partial v_i} - \frac{\partial H}{\partial x_i} \frac{\partial H}{\partial v_i} \right) \right] = 0. \end{aligned}$$

This implies that  $\rho(t, x, v) = \rho_b(x, v)$  satisfies (15) (observe that  $\frac{\partial}{\partial t} \rho(t, x, v) = 0$  because  $\rho(t, x, v)$  is constant in  $t$ ).

This shows that if  $x \sim \pi$ , then the solution  $x(\sigma)$  to (1) at any time  $\sigma$  is distributed according to  $\pi$ . This means that  $\pi$  is stationary for the Hamiltonian Markov Chain.

## References

- [1] Radford M. Neal. MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*, pages 47–95. Chapman and Hall/CRC, 2011.
- [2] Mark Girolami and Ben Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.
- [3] Lai, C.-H., Song, Y., Kim, D., Mitsufuji, Y., and Ermon, S. (2025). The principles of diffusion models. *arXiv preprint arXiv:2510.21890*.
- [4] Nathan E. Glatt-Holtz, Andrew J. Holbrook, Justin A. Krometis, Cecilia F. Mondaini, and Ami Sheth. Sacred and Profane: from the Involutive Theory of MCMC to Helpful Hamiltonian Hacks. *arXiv preprint arXiv:2410.17398*, 2024.