Lecture 5: Hamiltonian Monte Carlo

Tianpei Xie

Sep. 28th., 2022

Contents

1 Basics of Calculus of Variations

1.1 Fundamental theorems of Calculus of Variations

- See proof of following lemma from [Gelfand et al., 2000].

  Lemma 1.1 *If $\alpha(x)$ is continuous in $[a,b]$, and if*

  $$\int_a^b \alpha(x)h(x)dx = 0 \tag{1}$$

  for every continuous function $h(x) \in \mathcal{C}[a,b]$ such that $h(a) = h(b) = 0$, then $\alpha(x) = 0$ for all x in $[a,b]$

- **Lemma 1.2** *If $\alpha(x)$ is continuous in $[a,b]$, and if*

  $$\int_a^b \alpha(x)\dot{h}(x)dx = 0$$

  for every first-order differentiable function $h(x) \in \mathcal{D}_1[a,b]$ such that $h(a) = h(b) = 0$, then $\alpha(x) = c$ constant for all x in $[a,b]$

- **Lemma 1.3** *If $\alpha(x)$ is continuous in $[a,b]$, and if*

  $$\int_a^b \alpha(x)\ddot{h}(x)dx = 0$$

  for every second-order differentiable function $h(x) \in \mathcal{D}_2[a,b]$ such that $h(a) = h(b) = 0$ and $\dot{h}(a) = \dot{h}(b) = 0$, then $\alpha(x) = c_1\, x + c_0$ linear for all x in $[a,b]$ and c_0, c_1 are some constants.

- **Lemma 1.4** *If $\alpha(x), \beta(x)$ are continuous in $[a,b]$, and if*

  $$\int_a^b \left[\alpha(x)h(x) + \beta(x)\dot{h}(x)\right] dx = 0 \tag{2}$$

  for every first-order differentiable function $h(x) \in \mathcal{D}_1[a,b]$ such that $h(a) = h(b) = 0$, then $\beta(x)$ is differentiable, $\frac{d}{dx}\beta(x) = \alpha(x)$ for all x in $[a,b]$.

- **Definition** [Gelfand et al., 2000]
  Let $J[y]$ be a ***functional*** defined on some *normed linear space*, and let

  $$\Delta J[h] = J[y+h] - J[y]$$

  be its ***increment***, corresponding to the increment $h = h(x)$ of the "*independent variable*" $y = y(x)$. If y is fixed, $\Delta J[h]$ is a functional of h, in general a nonlinear functional. Suppose that

  $$\Delta J[h] = \phi[h] + \epsilon \|h\|$$

  where $\phi[h]$ is a linear functional and $\epsilon \to 0$ as $\|h\| \to 0$. Then the functional $J[h]$ is said to be ***differentiable***, and the *principal linear part* of the increment $\Delta J[h]$, i.e., the linear functional $\phi[h]$ which differs from $\Delta J[h]$ by an infinitesimal of order higher than 1 relative to $\|h\| \to 0$, is called the ***variation (or differential)*** of $J[y]$ and is denoted by $\delta J[h]$.

- **Theorem 1.5** ***(Necessary condition for extreme point of differentiable functional)*** *[Gelfand et al., 2000]*
  A ***necessary condition*** *for the differentiable functional $J[y]$ to have an extremum for $y = \hat{y}$ is that* ***its variation vanish*** *for $y = \hat{y}$, i.e., that*

$$\delta J[h] = 0 \tag{3}$$

  for $y = \hat{y}$ and all admissible h.

1.2 Euler-Lagrange equation

In the ***calculus of variations*** [Gelfand et al., 2000] and *classical mechanics*, the ***Euler-Lagrange equations*** are a system of second-order ordinary differential equations whose solutions are stationary points of the given action functional.

- **Problem formulation**: Let $F(x, y, z)$ be a function with continuous first and second (partial) derivatives with respect to all its arguments. Then, among all functions $y(x)$ which are continuously differentiable for $x \in [a, b]$ and satisfy the boundary conditions

$$y(a) = A,\ y(b) = B,$$

  find the function for which the functional

$$J[y] = \int_a^b F(x,\, y,\, \dot{y})dx \tag{4}$$

  has a ***weak extremum***.

- **Theorem 1.6** ***(Euler-Lagrange equations)*** *[Gelfand et al., 2000]*
  Let $J[y]$ be a functional of the forms:

$$J[y] = \int_a^b F(x,\, y,\, \dot{y})dx \tag{5}$$

  defined on the set of functions $y(x)$ which have continuous first derivatives in $[a, b]$ and satisfy the boundary conditions $y(a) = A, y(b) = B$. Then a ***necessary condition*** *for $J[y]$ to have an extremum for a given function $y(x)$ is that $y(x)$ satisfy* ***Euler's equation***

$$\frac{\partial F}{\partial y} - \frac{d}{dx}\frac{\partial F}{\partial \dot{y}} = 0. \tag{6}$$

  Proof: Suppose we give $y(x)$ an increment $h(x)$, where, in order for the function $y(x) + h(x)$ to continue to satisfy the boundary conditions, we must have $h(a) = h(b) = 0$.

  Then, since the corresponding increment of the functional ΔJ equals

$$\begin{aligned}\Delta J = J[y + h] - J[y] &= \int_a^b F(x,\, y + h,\, \dot{y} + \dot{h})dx - \int_a^b F(x,\, y,\, \dot{y})dx \\ &= \int_a^b \Big[F(x,\, y + h,\, \dot{y} + \dot{h}) - F(x,\, y,\, \dot{y})\Big]\, dx,\end{aligned}$$

  it follows by using Taylor's theorem that

$$\Delta J = \int_a^b \left[\frac{\partial F}{\partial y}(x, y, \dot{y})\, h + \frac{\partial F}{\partial \dot{y}}(x, y, \dot{y})\, \dot{h}\right] dx + \ldots,$$

where the dots denote terms of order higher than 1 relative to h and $\dot{h}$. The integral in the right-hand side represents the principal linear part of the increment $\Delta J[h]$, and hence the variation of $J[y]$ is

$$\delta J = \int_a^b \left[\frac{\partial F}{\partial y} h + \frac{\partial F}{\partial \dot{y}} \dot{h} \right] dx.$$

Using the necessary condition for extreme $y = y(x)$ of continous functional, we have

$$\delta J[h] = \int_a^b \left[\frac{\partial F}{\partial y} h + \frac{\partial F}{\partial \dot{y}} \dot{h} \right] dx = 0$$

for all admissible h. According to Lemma (2), $\alpha(x) = \frac{\partial F}{\partial y}$ and $\beta(x) = \frac{\partial F}{\partial \dot{y}}$,

$$\frac{\partial F}{\partial y} - \frac{d}{dx} \frac{\partial F}{\partial \dot{y}} = 0. \quad \blacksquare$$

- From geometry point of view, the equation (6) is a system of differential equations for the ***geodesic*** between $y(a)$ and $y(b)$.
- The equation (6) can be generalized to multiple variables. Let

$$J[\boldsymbol{x}] = J[x^1, \ldots, x^n] := \int_a^b \mathcal{L}(\boldsymbol{x}, \dot{\boldsymbol{x}}, t) dt = \int_a^b \mathcal{L}(x^1, \ldots, x^n, \dot{x}^1, \ldots, \dot{x}^n, t) dt \tag{7}$$

be the functional of Lagrangian function. The ***geodesic*** between $\boldsymbol{x}(a)$ and $\boldsymbol{x}(b)$ is computed by minimizing the functional in (7). From physics point of view, the integral evaluated along the extremal passing through two given points, i.e., two configurations of the system, is the "least action" corresponding to the motion of the system from the first configuration to the second.

Its solution is given by the ***Euler-Lagrange equation***

$$\frac{\partial \mathcal{L}}{\partial x^i} - \frac{d}{dt} \frac{\partial \mathcal{L}}{\partial \dot{x}^i} = 0, \forall\, i. \tag{8}$$

1.3 Hamilton's equations

- **Definition** The ***Hamiltonian (function)*** corresponding to the functional $J[\boldsymbol{x}]$ in (7) is defined as

$$\begin{aligned} \mathcal{H}(\boldsymbol{x}, \boldsymbol{v}, t) &:= \langle \boldsymbol{v}\,,\, \dot{\boldsymbol{x}} \rangle - \mathcal{L}(\boldsymbol{x}, \dot{\boldsymbol{x}}, t) \\ \mathcal{H}(x^1, \ldots, x^n, v_1, \ldots, v_n, t) &= \sum_i v_i \dot{x}^i - \mathcal{L}(x^1, \ldots, x^n, \dot{x}^1, \ldots, \dot{x}^n, t) \end{aligned} \tag{9}$$

- In this way, we can make a local transformation from the "variables" $(t, x^1, \ldots, x^n, \dot{x}^1, \ldots, \dot{x}^n, \mathcal{L})$ appearing in (7) to the new quantities $(t, x^1, \ldots, x^n, v_1, \ldots, v_n, \mathcal{H})$, called the ***canonical variables*** corresponding to the functional $J[\boldsymbol{x}] = J[x^1, \ldots, x^n]$.

- **Theorem 1.7** ***(Hamilton's equations or The canonical system of Euler's equations)*** *[Gelfand et al., 2000]*
Let $J[y]$ be a functional of the forms:

$$J[x^1, \ldots, x^n] = \int_a^b \mathcal{L}(x^1, \ldots, x^n, \dot{x}^1, \ldots, \dot{x}^n, t)dt \tag{10}$$

defined on the set of functions $x^i(t), i = 1, \ldots, n$ which have continuous first derivatives in $[a, b]$ and satisfy the boundary conditions $\boldsymbol{x}(a) = [x^1(a), \ldots, x^n(a)] = A$, $\boldsymbol{x}(b) = [x^1(b), \ldots, x^n(b)] = B$. The Hamiltonian $\mathcal{H}$ corresponding to functional J is defined as

$$\mathcal{H}(x^1, \ldots, x^n, v_1, \ldots, v_n, t) = \sum_i v_i \dot{x}^i - \mathcal{L}(x^1, \ldots, x^n, \dot{x}^1, \ldots, \dot{x}^n, t).$$

Then a ***necessary condition*** *for $J[x^1, \ldots, x^n]$ to have an extremum for a given function $\boldsymbol{x}(t) = [x^1(t), \ldots, x^n(t)]$ is that $\boldsymbol{x}(t)$ and $\boldsymbol{v}(t)$ satisfy* ***Hamilton's equations***

$$\frac{\partial \mathcal{H}}{\partial v_i} = \dot{x}^i, \quad \forall\, i \tag{11}$$

$$\frac{\partial \mathcal{H}}{\partial x^i} = -\frac{\partial \mathcal{L}}{\partial x^i} = -\dot{v}_i, \quad \forall\, i \tag{12}$$

$$\frac{\partial \mathcal{H}}{\partial t} = -\frac{\partial \mathcal{L}}{\partial t} \tag{13}$$

Proof: We can find the total differential of Lagrangian function

$$\begin{aligned}
d\mathcal{L} &= \sum_i \frac{\partial \mathcal{L}}{\partial x^i} dx^i + \sum_i \frac{\partial \mathcal{L}}{\partial \dot{x}^i} d\dot{x}^i + \frac{\partial \mathcal{L}}{\partial t} dt \\
&= \sum_i \frac{\partial \mathcal{L}}{\partial x^i} dx^i + \sum_i \left[d\left(\frac{\partial \mathcal{L}}{\partial \dot{x}^i} \dot{x}^i \right) - \dot{x}^i d\left(\frac{\partial \mathcal{L}}{\partial \dot{x}^i} \right) \right] + \frac{\partial \mathcal{L}}{\partial t} dt \\
0 &= \sum_i d\left(\frac{\partial \mathcal{L}}{\partial \dot{x}^i} \dot{x}^i \right) - d\mathcal{L} + \sum_i \frac{\partial \mathcal{L}}{\partial x^i} dx^i - \sum_i \dot{x}^i d\left(\frac{\partial \mathcal{L}}{\partial \dot{x}^i} \right) + \frac{\partial \mathcal{L}}{\partial t} dt \\
&= \sum_i d\left(\frac{\partial \mathcal{L}}{\partial \dot{x}^i} \dot{x}^i - \mathcal{L} \right) + \sum_i \left[\frac{\partial \mathcal{L}}{\partial x^i} dx^i - \dot{x}^i d\left(\frac{\partial \mathcal{L}}{\partial \dot{x}^i} \right) \right] + \frac{\partial \mathcal{L}}{\partial t} dt \\
d\left(\sum_i \frac{\partial \mathcal{L}}{\partial \dot{x}^i} \dot{x}^i - \mathcal{L} \right) &= \sum_i \left[-\frac{\partial \mathcal{L}}{\partial x^i} dx^i + \dot{x}^i d\left(\frac{\partial \mathcal{L}}{\partial \dot{x}^i} \right) \right] - \frac{\partial \mathcal{L}}{\partial t} dt \\
d\left(\sum_i v_i \dot{x}^i - \mathcal{L} \right) &= \sum_i \left(-\frac{\partial \mathcal{L}}{\partial x^i} \right) dx^i + \sum_i \dot{x}^i dv_i - \frac{\partial \mathcal{L}}{\partial t} dt \quad \text{let } v_i := \frac{\partial \mathcal{L}}{\partial \dot{x}^i} \\
d\mathcal{H} &= \sum_i \left(-\frac{\partial \mathcal{L}}{\partial x^i} \right) dx^i + \sum_i \dot{x}^i dv_i - \frac{\partial \mathcal{L}}{\partial t} dt \\
\text{Also } d\mathcal{H} &= \sum_i \frac{\partial \mathcal{H}}{\partial x^i} dx^i + \sum_i \frac{\partial \mathcal{H}}{\partial v_i} dv_i + \frac{\partial \mathcal{H}}{\partial t} dt
\end{aligned}$$

Therefore we have

$$\frac{\partial \mathcal{H}}{\partial x^i} = -\frac{\partial \mathcal{L}}{\partial x^i} = -\frac{d}{dt}\frac{\partial \mathcal{L}}{\partial \dot{x}} = -\dot{v}_i \quad \text{(by Euler-Lagrange equation)}$$

$$\frac{\partial \mathcal{H}}{\partial v_i} = \dot{x}^i$$
$$\frac{\partial \mathcal{H}}{\partial t} = -\frac{\partial \mathcal{L}}{\partial t} \quad \blacksquare$$

- In the case of time-independent $\mathcal{H}$ and $\mathcal{L}$, i.e. $\frac{\partial \mathcal{H}}{\partial t} = -\frac{\partial \mathcal{L}}{\partial t} = 0$, the ***Hamiltonian dynamic*** is defined by $2n$ first-order differential equations:

$$\frac{dx^i}{dt} = \frac{\partial \mathcal{H}}{\partial v_i}, \quad i = 1, \ldots, n$$
$$\frac{dv_i}{dt} = -\frac{\partial \mathcal{H}}{\partial x^i}, \quad i = 1, \ldots, n \tag{14}$$
$$\Rightarrow \frac{d\boldsymbol{z}}{dt} = \boldsymbol{\Omega} \nabla_{\boldsymbol{z}} \mathcal{H} \tag{15}$$

  where $\boldsymbol{z} := (\boldsymbol{x}, \boldsymbol{v})$ and $\boldsymbol{\Omega} = \begin{bmatrix} \mathbf{0} & \boldsymbol{I}_n \\ -\boldsymbol{I}_n & \mathbf{0} \end{bmatrix}$. These differential equations describes the dynamic of a particle moving along the **geodesic curve** from ***position*** $\boldsymbol{x}(a)$ with ***velocity*** (or ***momentum***) $\boldsymbol{v}(a)$ (i.e. configuration $(\boldsymbol{x}(a), \boldsymbol{v}(a))$) to position $\boldsymbol{x}(b)$ with $\boldsymbol{v}(b)$ (i.e. configuration $(\boldsymbol{x}(b), \boldsymbol{v}(b))$). The vector $\boldsymbol{v}$ defines a ***vector fields*** that is *tangent* to the *potential* $\mathcal{L}$ at position $\boldsymbol{x}$.

- An alternative representation for Hamiltonian dynamic is called ***variational equations*** [Leimkuhler and Reich, 2004]:

$$\frac{d}{dt}\boldsymbol{\xi} = \boldsymbol{\Omega} \nabla^2_{\boldsymbol{z},\boldsymbol{z}} \mathcal{H}(\boldsymbol{z}_t)\, \boldsymbol{\xi}. \tag{16}$$

  where $\boldsymbol{\xi}(t) = \partial_{\boldsymbol{z}} \phi_t(\boldsymbol{z}_0)\, \delta \boldsymbol{z}_0$ for Hamiltonian flow map ϕ_t.

- The definition of Hamiltonian can be formulated as the sum of ***kinetic energy*** $\mathcal{K}$ and ***potential energy*** $\mathcal{V}$ (i.e. *frictionless* motion).

- The Hamiltonian dynamic captures the **geometry** of surface by specifying the vector field $\boldsymbol{F} = (\frac{d\boldsymbol{x}}{dt}, \frac{d\boldsymbol{v}}{dt})$ on the surface.

1.4 Properties of Hamiltonian dynamics

Let the ***Hamiltonian flow*** $\phi_t : (\boldsymbol{x}, \boldsymbol{v}) \mapsto (\boldsymbol{x}_t(\boldsymbol{x}, \boldsymbol{v}), \boldsymbol{v}_t(\boldsymbol{x}, \boldsymbol{v}))$ be the position and velocity after time t starting from $(\boldsymbol{x}, \boldsymbol{v})$. The Hamiltonian flow satisfies following ***properties***:

1. ***Reversibility***: Hamiltonian dynamics of the form mentioned above are ***time reversible*** for $t \geq 0$:

$$\phi_t(\boldsymbol{x}_t(\boldsymbol{x}, \boldsymbol{v}), -\boldsymbol{v}_t(\boldsymbol{x}, \boldsymbol{v})) = (\boldsymbol{x}, -\boldsymbol{v}). \tag{17}$$

   The mapping $\phi_t : (\boldsymbol{x}, \boldsymbol{v}) \mapsto (\boldsymbol{x}_t(\boldsymbol{x}, \boldsymbol{v}), \boldsymbol{v}_t(\boldsymbol{x}, \boldsymbol{v}))$ is one-to-one and has inverse mapping.

2. ***Conservation of the Hamiltonian***: A second property of the dynamics is that it keeps the ***Hamiltonian invariant*** (i.e. *conserved*).

$$\frac{d\mathcal{H}}{dt} = \sum_{i=1}^{n} \left[\frac{dx^i}{dt} \frac{\partial \mathcal{H}}{\partial x^i} + \frac{dv_i}{dt} \frac{\partial \mathcal{H}}{\partial v_i} \right]$$

$$= \sum_{i=1}^{n} \left[\frac{dx^i}{dt} \frac{dv_i}{dt} - \frac{dv_i}{dt} \frac{dx^i}{dt} \right] = 0 \tag{18}$$

To see this note that, since $\mathcal{H}$ does not depend on t explicitly (or $\frac{\partial \mathcal{H}}{\partial t} = 0$).

3. ***Volume Preservation***: A third fundamental property of Hamiltonian dynamics is that it ***preserves volume in phase space*** $(\boldsymbol{x}, \boldsymbol{v})$ space. Formally, let $\boldsymbol{F} = (\frac{d\boldsymbol{x}}{dt}, \frac{d\boldsymbol{v}}{dt})$ be the vector field associated to the Hamiltonian in the phase space $\mathbb{R}^n \times \mathbb{R}^n$. First note that the ***divergence of $\boldsymbol{F}$ is zero***:

$$\begin{aligned} \nabla \cdot \boldsymbol{F} = \operatorname{div} \boldsymbol{F} &= \sum_{i=1}^{n} \left[\frac{\partial}{\partial x^i} \frac{dx^i}{dt} + \frac{\partial}{\partial v_i} \frac{dv_i}{dt} \right] \\ &= \sum_{i=1}^{n} \left[\frac{\partial}{\partial x^i} \frac{\partial \mathcal{H}}{\partial v_i} - \frac{\partial}{\partial v_i} \frac{\partial \mathcal{H}}{\partial x^i} \right] = 0 \end{aligned} \tag{19}$$

Since divergence represents the volume density of the outward flux of a vector field from an infinitesimal volume around a given point, it being zero everywhere implies volume preservation.

Another way to see this is that divergence is the ***trace*** of the ***Jacobian*** of the map $\boldsymbol{F}$, and the trace of the Jacobian is the ***derivative*** of the ***determinant of the Jacobian***. Hence, the trace being 0 implies that the determinant of the Jacobian of $\boldsymbol{F}$ does not change.

4. ***Symplecticness***: Volume preservation is also a consequence of Hamiltonian dynamics being ***symplectic***. Let $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{v})$, and define $\boldsymbol{\Omega}$ as

$$\boldsymbol{\Omega} = \begin{bmatrix} \mathbf{0} & \boldsymbol{I}_n \\ -\boldsymbol{I}_n & \mathbf{0} \end{bmatrix}.$$

Note $\boldsymbol{\Omega}^{-1} = \boldsymbol{\Omega}^T = -\boldsymbol{\Omega}$.

The ***symplecticness condition*** is that the Jacobian matrix, $(\partial_{\boldsymbol{z}} \boldsymbol{\phi}_t)$ of the mapping ϕ_t satisfies

$$(\partial_{\boldsymbol{z}} \boldsymbol{\phi}_t)^T \boldsymbol{\Omega}^{-1} (\partial_{\boldsymbol{z}} \boldsymbol{\phi}_t) = \boldsymbol{\Omega}^{-1} \tag{20}$$

This implies volume conservation, since $\det (\partial_{\boldsymbol{z}} \boldsymbol{\phi}_t)^T \det \left(\boldsymbol{\Omega}^{-1}\right) \det (\partial_{\boldsymbol{z}} \boldsymbol{\phi}_t) = \det \left(\boldsymbol{\Omega}^{-1}\right)$ implies that $\det (\partial_{\boldsymbol{z}} \boldsymbol{\phi}_t)$ is one. When $n > 1$, the ***symplecticness condition is stronger than volume preservation***. Hamiltonian dynamics and the symplecticness condition can be generalized to where $\boldsymbol{\Omega}$ is any matrix for which $\boldsymbol{\Omega}^T = \boldsymbol{\Omega}^{-1}$ and $\det \boldsymbol{\Omega} \neq 0$.

Crucially, reversibility, preservation of volume, and symplecticness can be maintained exactly even when, as is necessary in practice, Hamiltonian dynamics is approximated.

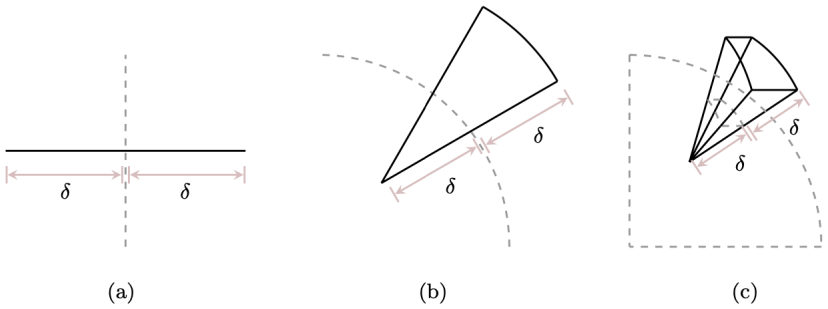

FIG 2. *The dominance of volume away from any point in parameter space can also be seen from a spherical perspective, where we consider the volume contained radial distance δ both interior to and exterior to a D-dimensional spherical shell, shown here with dashed lines. (a) In one dimension the spherical shell is a line and volumes interior and exterior are equivalent. (b) In two dimensions the spherical shell becomes circle and there is more volume immediately outside the shell than immediately inside. (c) The exterior volume grows even larger relative to the interior volume in three dimensions, where the spherical shell is now a the surface of a sphere. In fact, with increasing dimension the exterior volume grows exponentially large relative to the interior volume, and very quickly the volume around the mode is dwarfed by the volume away from the mode.*

Figure 1: The concentration of measure around the shell of a unit ball [Betancourt, 2017]

2 Markov Chain Monte Carlo from Hamiltonian Dynamics

2.1 Monte Carlo in high dimensional spaces

- One of the characteristic properties of **high-dimensional spaces** is that there is *much more volume outside any given neighborhood than inside of it*. Figure 1 shows that as the dimensionality increases, more vol lies in the shell of the ball not in the core.
- On the other hand, the complimentary neighborhood ***far away from the mode*** features a *much larger volume*, but the ***vanishing densities*** lead to similarly negligible contributions expectations. The only significant contributions come from the neighborhood between these two extremes known as ***the typical set***. (Figure 2) The Importantly, because probability densities and volumes transform oppositely under any ***reparameterization***, the typical set is an invariant object that does not depend on the irrelevant details of any particular choice of parameters.

  As the dimension of parameter space increases, however, the tension between the density and the volume grows and the typical set becomes *more singular*.
- The **success** of Markov Chain Monte Carlo lies in its ability to explore the parameter space while preserving the target distribution. The resulting Markov chain will drift into and then across the typical set regardless of its initial state, providing a powerful quantification of the typical set from which we can derive accurate expectation estimators.
- Under ideal conditions, Markov chains explore the target distribution in **three distinct phases** [Betancourt, 2017]:

  1. In the first phase the Markov chain *converges towards* the typical set from its initial position in parameter space while the Markov chain Monte Carlo estimators suffer from ***strong biases***. Consequently, it is common practice to ***warm up*** the Markov chain by

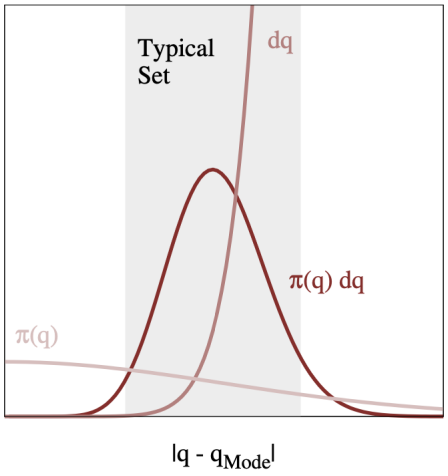

FIG 3. *In high dimensions a probability density, $\pi(q)$, will concentrate around its mode, but the volume over which we integrate that density, $\mathrm{d}q$, is much larger away from the mode. Contributions to any expectation are determined by the product of density and volume, $\pi(q)\,\mathrm{d}q$, which then concentrates in a nearly-singular neighborhood called the typical set (grey).*

Figure 2: The typical set of $\pi(q)dq$ is not close to the mode nor far away from it. [Betancourt, 2017]

throwing away those initial converging samples before computing Markov chain Monte Carlo estimators [Liu, 2001].

2. The second phase begins once the Markov chain finds the typical set and persists through the first sojourn across the typical set. This initial exploration is extremely effective and the accuracy of Markov chain Monte Carlo estimators rapidly improves as the bias from the initial samples is eliminated.

3. The third phase consists of all subsequent exploration where the Markov chain refines its exploration of the typical set and the precision of the Markov chain Monte Carlo estimators improves, albeit at a slower rate. In this phase, the central limit theorem holds.

- Random Walk Metropolis-Hastings is **doomed to fail** in high dimensional space due to its inefficiency and low acceptance rate. There are an exponential number of directions in which to guess but only a singular number of directions that stay within the typical set and pass the check. (Figure 3) Regardless of how we tune the covariance of the Random Walk Metropolis proposal or the particular details of the target distribution, the resulting Markov chain will explore the typical set extremely slowly in all but the lowest dimensional spaces.

- In order to make large jumps away from the initial point, and into new, unexplored regions of the typical set, we need to exploit information about the ***geometry*** of the *typical set* itself. Specifically, we need transitions that can follow those ***contours of high probability mass***, coherently gliding through the typical set. (Figure 4)

- Hamiltonian Monte Carlo is the unique procedure for automatically generating this coherent exploration for sufficiently well-behaved target distributions.

FIG 10. *In high dimensions, the Random Walk Metropolis proposal density (green) is strongly biased towards the outside of the typical set where the target density, and hence the Metropolis acceptance probability vanishes. (a) If the proposal variances are large then the proposals will stray too far away from the typical set and are rejected. (b) Smaller proposal variances stay within the typical set and hence are accepted, but the resulting transition density concentrates tightly around the initial point. Either way we end up with a Markov chain that explores the typical set very, very slowly.*

Figure 3: The challenge of random walk Metropolis-Hastings on high dimensional space. [Betancourt, 2017]

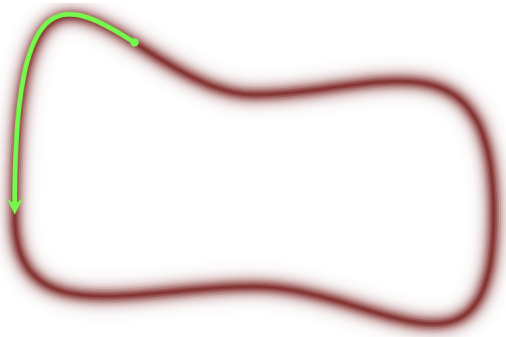

FIG 11. *Most Markov transitions are diffusive, concentrating around the initial point such that the corresponding Markov chains linger in small neighborhoods of the typical set for long periods of time. In order to maximize the utility of our computational resources we need* coherent *Markov transitions that are able to glide across the typical set towards new, unexplored neighborhoods.*

Figure 4: The Markov chain that is able to explore the contours of high probability mass is needed.

2.2 Hamiltonian Monte Carlo

- Compare to MCMC, ***Hamiltonian Monte Carlo (HMC)*** [Brooks et al., 2011] has some unique characteristics:
  - HMC reply on the ***Hamiltonian dynamic*** to **explore** the parameter space. As oppose to the *stochastic process* from MCMC, the HMC exploration process is ***deterministic*** based on the *differential equations* (11), (12). HMC only need to simulate the **initial velocity/momentum** $\boldsymbol{v}(0)$ to start the Hamiltonian dynamic.
  - HMC ***preserves the target distribution*** π by *properties of Hamiltonian dynamic*, while the MCMC preserve the target distribution as the *stationary distribution* of Markov chain.
  - Unlike the *diffusion transition kernel* in MCMC, the transition process of HMC encodes ***geometrical information*** of the typical set via its associated ***vector field***. In other words, instead of fumbling around parameter space with random, uninformed jumps, we can follow the direction assigned to each at point for a small distance. Continuing this process traces out a coherent trajectory through the typical set that efficiently moves us far away from the initial point to new, unexplored regions of the typical set *as quickly as possible*.
  - The performance of MCMC and its special case, Gibbs sampling, depend on a particular ***parameterization*** of the target distribution. On the other hand, the design of Hamiltonian dynamic carefully remove the dependency on parameterization with additional geometric constraints while twisting the directions to align with the typical set.
  - Compared to Random-walk Metropolis-Hastings, HMC reduces the correlation between successive sampled states by proposing moves to distant states which maintain a high probability of acceptance.
- Inspired by classical mechanic systems, the **key idea** of ***Hamiltonian Monte Carlo (HMC)*** is introduce ***auxiliary momentum parameters*** in order to twist the gradient vector field into a vector field aligned with the typical set, and hence one capable of generating efficient exploration.
- Consider the target distribution as

$$\pi(\boldsymbol{x}, \boldsymbol{v}) \propto \frac{1}{Z} \exp\left(-\mathcal{H}(\boldsymbol{x}, \boldsymbol{v})\right), \tag{21}$$

  where $\boldsymbol{v}$ is the ***auxiliary momentum parameters***, $\boldsymbol{z} := (\boldsymbol{x}, \boldsymbol{v})$ is called ***phase-space parameterization***. The joint probability $\pi(\boldsymbol{x}, \boldsymbol{v})$ distribution on phase space is called the ***canonical distribution*** or phase-space distribution.

  Then we can factorize π into

$$\pi(\boldsymbol{x}, \boldsymbol{v}) = \pi(\boldsymbol{v}|\boldsymbol{x})\pi(\boldsymbol{x}) \tag{22}$$

$$\begin{aligned} -\log \pi(\boldsymbol{x}, \boldsymbol{v}) &= -\log \pi(\boldsymbol{v}|\boldsymbol{x}) - \log \pi(\boldsymbol{x}) \\ &= \mathcal{K}(\boldsymbol{x}, \boldsymbol{v}) + \mathcal{V}(\boldsymbol{x}) := \mathcal{H}(\boldsymbol{x}, \boldsymbol{v}). \end{aligned} \tag{23}$$

  where $\mathcal{K}(\boldsymbol{x}, \boldsymbol{v})$ is the *kinetic energy* and $\mathcal{V}(\boldsymbol{x})$ is the *potential energy*. The potential energy is completely determined by the target distribution while the kinetic energy is ***unconstrained*** and must be specified by the implementation.

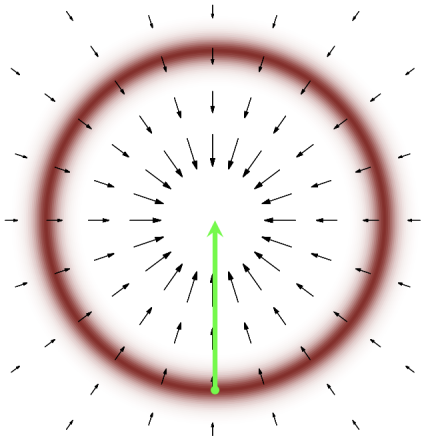

FIG 13. *The gradient of the target probability density function encodes information about the geometry of the typical set, but not enough to guide us through the typical set by itself. Following along the gradient instead pulls us away from the typical set and towards the mode of the target density. In order to generate motion through the typical set we need to introduce additional structure that carefully twists the gradient into alignment with the typical set.*

Figure 5: The vector fields encode geometrical information and the gradient pull towards the mode of distribution.

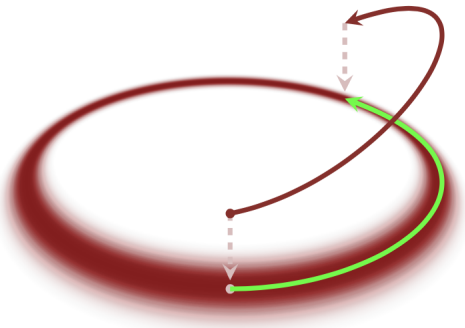

FIG 19. *By constructing a probability distribution on phase space that marginalizes to the target distribution, we ensure that the typical set on phase space projects to the typical set of the target distribution. In particular, if we can construct trajectories that efficiently explore the joint distribution (black) they will project to trajectories that efficiently explore the target distribution (green).*

Figure 6: By constructing a probability distribution on phase space that marginalizes to the target distribution, we ensure that the typical set on phase space projects to the typical set of the target distribution. [Betancourt, 2017]

This factorization guarantees that any trajectories exploring the typical set of the phase space distribution $\pi(\boldsymbol{x}, \boldsymbol{v})$ will project to trajectories exploring the typical set of the target distribution $\pi(\boldsymbol{x})$.

- Note that $\mathcal{H}(\boldsymbol{x}, \boldsymbol{v})$ is ***independent*** of the details of any ***parameterization***, so does the *canonical density* $\pi(\boldsymbol{x}, \boldsymbol{v})$ in (21). Moreover, $\mathcal{H}(\boldsymbol{x}, \boldsymbol{v})$ captures the ***invariant probabilistic structure*** of the phase-space distribution and, most importantly, the ***geometry*** of its typical set.

- Given (23), recall from (14) the Hamilton's equations:

$$\begin{aligned} \frac{dx^i}{dt} &= \frac{\partial \mathcal{K}}{\partial v_i} = -\frac{\partial \log \pi(\boldsymbol{v}|\boldsymbol{x})}{\partial v_i}, \quad i = 1, \ldots, d \\ \frac{dv_i}{dt} &= -\frac{\partial \mathcal{K}}{\partial x^i} - \frac{\partial \mathcal{V}}{\partial x^i} = \frac{\partial \log \pi(\boldsymbol{v}|\boldsymbol{x})}{\partial x^i} + \frac{\partial \log \pi(\boldsymbol{x})}{\partial x^i}, \quad i = 1, \ldots, d \end{aligned} \tag{24}$$

2.3 Idealized Hamiltonian Monte Carlo

- Let $\phi_t : (\boldsymbol{x}, \boldsymbol{v}) \mapsto (\boldsymbol{x}_t(\boldsymbol{x}, \boldsymbol{v}), \boldsymbol{v}_t(\boldsymbol{x}, \boldsymbol{v}))$ be the trajectory characterized by differential equations (24). $\phi_t(\boldsymbol{x}, \boldsymbol{v})$ is the position and velocity/momentum at time t starting from $(\boldsymbol{x}, \boldsymbol{v})$.

- The ***idealized Hamiltonian Monte Carlo*** [Betancourt, 2017, Vishnoi, 2021] is described as below:

  1. For $t = 1, 2, \ldots, k$:

     (a) Given $\boldsymbol{X}_{t-1}$, generate a momentum $\boldsymbol{V}_{t-1}$ from conditional distribution $\pi(\boldsymbol{v}|\boldsymbol{X}_{t-1})$;

     (b) Set $(\boldsymbol{X}_t, \boldsymbol{V}_t) = \phi_T(\boldsymbol{X}_{t-1}, \boldsymbol{V}_{t-1})$ by integrating Hamilton's equations for some time T

  2. Return $\boldsymbol{X}_k$ after projecting away the momentum.

  Composing these steps together yields a ***Hamiltonian Markov transition*** composed of *random trajectories* that rapidly explore the target distribution.

- The sequence $(\boldsymbol{X}_t, \boldsymbol{V}_t)$ is a **Markov chain** since $(\boldsymbol{X}_{t+1}, \boldsymbol{V}_{t+1})$ only depends on the initial condition of Hamiltonian dynamic $(\boldsymbol{X}_t, \boldsymbol{V}_t)$. Consequently, the sub-chain $(\boldsymbol{X}_t)$ is also a Markov chain.

- Note that since the Hamiltonian dynamic remains in typical set throughout iterations, there is no need for rejection, i.e. the acceptance rate is 1. Given T large enough, a new sample $\boldsymbol{X}_{t+1}$ would have less correlation to $\boldsymbol{X}_t$. Thus it is expected that

2.4 The Natural Geometry of Phase Space

- One of the characteristic properties of Hamilton's equations is that they *conserve the value of the Hamiltonian*. In other words, every Hamiltonian trajectory is confined to an ***energy level set***,

$$\mathcal{H}^{-1}(E) = \{(\boldsymbol{x}, \boldsymbol{v}) : \mathcal{H}(\boldsymbol{x}, \boldsymbol{v}) = E\},$$

  which, save for some ignorable exceptions, are all $(2\,d - 1)$-dimensional, ***compact*** surfaces in phase space. In fact, once we've removed any singular level sets, the entirety of phase space

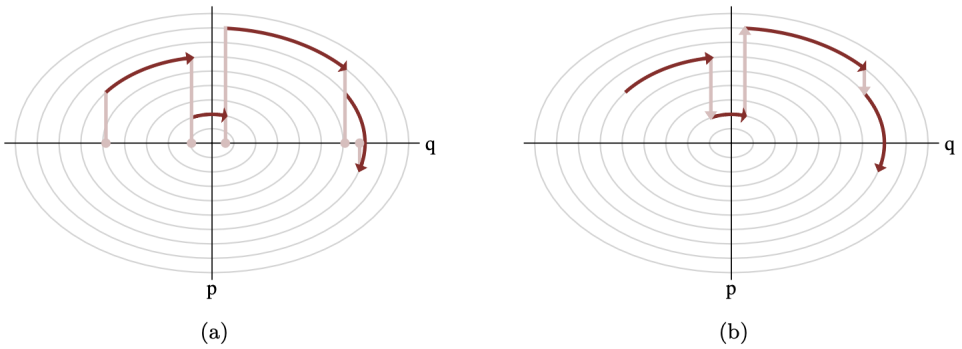

FIG 22. *(a) Each Hamiltonian Markov transition lifts the initial state onto a random level set of the Hamiltonian, which can then be explored with a Hamiltonian trajectory before projecting back down to the target parameter space. (b) If we consider the projection and random lift steps as a single momentum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

Figure 7: Using level-sets, we see that HMC first explore within level set and then jump from one level set to another via random walk. [Betancourt, 2017]

neatly decomposes, or ***foliates*** into ***concentric level sets***. Consequently, we can specify any point in phase space by first specifying the energy of the level set it falls on, E, and the position within that level set, $\boldsymbol{\theta}_E$.

- Correspondingly the canonical distribution on phase space admits a ***microcanonical decomposition***,

$$\pi(\boldsymbol{x}, \boldsymbol{v}) = \pi(\boldsymbol{\theta}_E \,|\, E)\pi(E),$$

across this ***foliation***. The conditional distribution over each level set, $\pi(\boldsymbol{\theta}_E \,|\, E)$, is called the ***microcanonical distribution***, while the distribution across the level sets, $\pi(E)$, is called the ***marginal energy distribution***.

- This microcanonical decomposition is particularly well-suited to analyzing the Hamiltonian transition. To see this more clearly, consider a Hamiltonian Markov chain consisting of multiple transitions. Each Hamiltonian trajectory explores a level set while the intermediate *projections* and *lifts* define a random jump between the level sets themselves. Consequently, the entire Hamiltonian Markov chain decouples into ***two distinct phases***:

  1. ***deterministic*** exploration of individual *level sets*
  2. a ***stochastic*** exploration ***between*** the level sets themselves

2.5 Property of Hamiltonian Monte Carlo

- Let μ be the Lebesgue measure on $\mathbb{R}^d \times \mathbb{R}^d$ with respect to which all densities are defined.

  Theorem 2.1 *(HMC Preserves the Target Density) [Vishnoi, 2021] Suppose* $(\boldsymbol{X}, \boldsymbol{V})$ *is a sample from the density*

$$\pi(\boldsymbol{x}, \boldsymbol{v}) = \frac{1}{Z} \exp\left(-\mathcal{H}(\boldsymbol{x}, \boldsymbol{v})\right) d\mu(\boldsymbol{x}, \boldsymbol{v})$$

where $Z = \int \exp\left(-\mathcal{H}(\boldsymbol{x}, \boldsymbol{v})\right) d\mu(\boldsymbol{x}, \boldsymbol{v})$ *is the partition function. Let* $T > 0$ *be the step size of the HMC. Then the density of* $\phi_T(\boldsymbol{X}, \boldsymbol{V})$ *is* π *for any* $T \geq 0$*. Moreover the density of* $\phi_T(\boldsymbol{X}_{t-1}, \boldsymbol{V}_{t-1})$*, where* $\boldsymbol{V}_{t-1} \sim \pi(\boldsymbol{v}|\boldsymbol{X}_{t-1})$ *is also* π*. Thus, the idealized HMC algorithms preserves* π*.*

Proof: For $T \geq 0$, let $(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{v}}) = \phi_T(\boldsymbol{x}, \boldsymbol{v})$. Then, it follows from the *preservation of Hamiltonian along trajectories* that

$$\mathcal{H}(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{v}}) = \mathcal{H}(\phi_T(\boldsymbol{x}, \boldsymbol{v})) = \mathcal{H}(\boldsymbol{x}, \boldsymbol{v}).$$

Thus the function $\exp(-\mathcal{H}(\boldsymbol{x}, \boldsymbol{v})) = \exp(-\mathcal{H}(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{v}}))$

Let μ^* be the *pushforward* of μ under the map ϕ_T, i.e. $(\phi_T)_{\#}\mu = \mu^*$. The property that Hamiltonian dynamics *preserves volume* in phase space states that the determinant of the Jacobian of the map ϕ_T, $\det\left(\partial_{\boldsymbol{z}} \boldsymbol{\phi}_T\right) = 1$. Therefore,

$$d\mu^* = \det\left(\partial_{\boldsymbol{z}} \boldsymbol{\phi}_T\right) \, d\mu = d\mu,$$

i.e. $\mu^* = \mu$. Thus, the normalization factor

$$Z^* = \int \exp(-\mathcal{H}(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{v}})) d\mu^* = \int \exp(-\mathcal{H}(\boldsymbol{x}, \boldsymbol{v})) d\mu = Z.$$

Therefore the target distribution π remains invariant under ϕ_T. ∎

2.6 Euclidean-Gaussian Kinetic Energies

- The first substantial ***degree of freedom*** in the Hamiltonian Monte Carlo method that we can tune is the choice of the conditional probability distribution over the momentum or, equivalently, **the choice of a kinetic energy function**. Along with the target distribution, this choice completes the probabilistic structure on phase space which then determines the geometry of the microcanonical decomposition.
- The simplest choice is to allow velocity/momentum $\boldsymbol{V}_t$ being ***independent*** of $\boldsymbol{X}_t$, i.e. $\pi(\boldsymbol{v}|\boldsymbol{x}) := g(\boldsymbol{v})$. This is equivalent to assuming that *the phase-space geometry* is ***flat***. The Hamiltonian Monte Carlo corresponds to a *random-walk* across a foliation of the phase-space.
- A natural choice of $g(\boldsymbol{v})$ is Normal distribution $\mathcal{N}(\boldsymbol{v} \,|\, \boldsymbol{0}, \boldsymbol{\Sigma}_v)$. Here, the covariance matrix $\boldsymbol{\Sigma}_v$ encodes an Euclidean metric in the phase-space. The corresponding kinetic energy is a ***Euclidean-Gaussian kinetic energy***

$$\mathcal{K}(\boldsymbol{v}) = \frac{1}{2} \boldsymbol{v}^T \boldsymbol{\Sigma}_v^{-1} \boldsymbol{v} + \log\det \boldsymbol{\Sigma}_v + const.. \tag{25}$$

And the Hamilton's equation is simplified as

$$\begin{aligned} \frac{d\boldsymbol{x}}{dt} &= \nabla_{\boldsymbol{v}} \mathcal{K} = \boldsymbol{\Sigma}_v^{-1} \boldsymbol{v} \\ \frac{d\boldsymbol{v}}{dt} &= -\nabla_{\boldsymbol{x}} \mathcal{V} = \nabla_{\boldsymbol{x}} E \end{aligned} \tag{26}$$

where the target distribution $\pi(\boldsymbol{x}) \propto \exp(-E(\boldsymbol{x}))$.

In the physical perspective the Euclidean metric $\boldsymbol{\Sigma}_v$ is known as the ***mass matrix***, a term that has consequently become common in the Hamiltonian Monte Carlo literature.

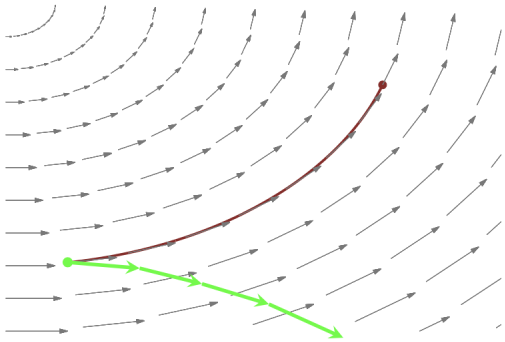

FIG 27. *The approximate solutions of most numerical integrators tend to drift away from the exact solutions. As the system is integrated longer and longer, errors add coherently and push the numerical trajectory away from the exact trajectory.*

Figure 8: Most numerical methods for PDE does not preserve volume. [Betancourt, 2017]

- Because the Euclidean structure over the momentum is dual to the Euclidean structure over the parameters, its interactions with the target distribution are straightforward to derive. Applying the transformation $\boldsymbol{p}' = \boldsymbol{\Sigma}_v^{-1/2} \boldsymbol{v}$ simplifies the kinetic energy, but remember that we have to apply the ***opposite transformation*** to the parameters, $\boldsymbol{x}' = \boldsymbol{\Sigma}_v^{1/2} \boldsymbol{x}$, to preserve the Hamiltonian geometry. Consequently, a choice of $\boldsymbol{\Sigma}_v^{-1}$ effectively ***rotates*** and then ***rescales*** the target **parameter** space, potentially ***correlating*** or ***de-correlating*** the target distribution and correspondingly ***warping*** the energy level sets.

  In particular, as the inverse Euclidean metric more closely resembles the covariance of the target distribution it de-correlates the target distribution, resulting in energy level sets that are more and more uniform and hence easier to explore.

- Thus the ***optimal choice*** of Euclidean-Gaussian kinetic energy is the *inverse* of covariance on $\boldsymbol{x}$:

$$\boldsymbol{\Sigma}_v^{-1} = \mathbb{E}_\pi \left[(\boldsymbol{X} - \boldsymbol{\mu}_X) (\boldsymbol{X} - \boldsymbol{\mu}_X)^T \right]$$

3 Hamiltonian Monte Carlo in Practice

- The ***main obstruction*** to implementing the Hamiltonian Monte Carlo method is **generating** the Hamiltonian **trajectories** themselves. Aside from a few trivial examples, we *cannot solve Hamilton's equations exactly* and any implementation must instead solve them numerically. **Numerical inaccuracies**, however, can quickly compromise the utility of even the most well-tuned Hamiltonian transition.

  The more accurately we can numerically solve the system of ODEs, the more effective our implementation will be.

- Common ODE solvers suffer from an issue of ***drift***. As we numerically solve longer and longer trajectories the error in the solvers adds coherently, pushing the approximate trajectory away from the true trajectory and the typical set that we want to explore (Figure 8). Moreover, the magnitude of this drift rapidly increases with the dimension of phase space. Thus these solvers are limited to solve short Hamiltonian trajectories, which is inefficiently explore the energy level sets.

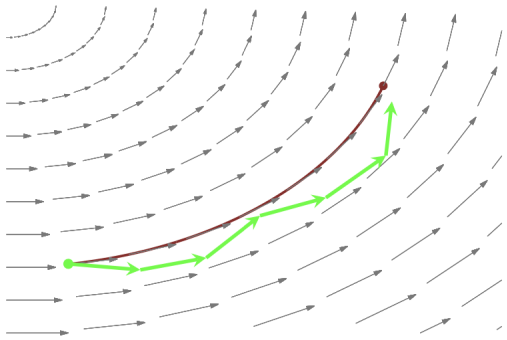

FIG 28. *Symplectic integrators generate numerical trajectories that are incompressible like the exact Hamiltonian trajectory they approximate. Consequently their approximation error cannot add up coherently to pull the numerical trajectories away from the exact trajectories. Instead the numerical trajectories oscillate around the exact level set, even as we integrate for longer and longer times.*

Figure 9: Symplectic integrator will oscillate near exact level set even for long integration times. [Betancourt, 2017]

- Fortunately, we can use the geometry of phase space itself to construct an extremely powerful family of numerical solvers, known as ***symplectic integrators*** [Liu, 2001, Leimkuhler and Reich, 2004, Haier et al., 2006, Betancourt, 2017, Vishnoi, 2021], that are robust to phenomena like drift and enable high-performance implementations of the Hamiltonian Monte Carlo method.

3.1 Symplectic Integrators

- Symplectic integrators are powerful because the numerical trajectories they generate exactly ***preserve phase space volume***, just like the Hamiltonian trajectories they are approximating. Consequently, the numerical trajectories cannot drift away from the exact energy level set, instead oscillating near it even for long integration times.
- For independent momentum distribution like the Euclidean-Gaussian kinetic energy, the symplectic integrator is called ***leapfrog integrator*** [Brooks et al., 2011].
- The ***Leapfrog Integrator*** is described as below, where ϵ is the time discretization, or step size:

  1. Initialization: $\boldsymbol{x}_0 \leftarrow \boldsymbol{x}$, $\boldsymbol{v}_0 \leftarrow \boldsymbol{v}$.
  2. For $0 \leq t \leq \lfloor \frac{T}{\epsilon} \rfloor$:

  $$(a) \quad \boldsymbol{v}_{t+\frac{1}{2}} \leftarrow \boldsymbol{v}_t - \frac{\epsilon}{2} \nabla_{\boldsymbol{x}} \mathcal{V}(\boldsymbol{x}_t)$$
  $$(b) \quad \boldsymbol{x}_{t+1} \leftarrow \boldsymbol{x}_t + \epsilon \, \boldsymbol{\Sigma}_v^{-1} \boldsymbol{v}_{t+\frac{1}{2}}$$
  $$(c) \quad \boldsymbol{v}_{t+1} \leftarrow \boldsymbol{v}_{t+\frac{1}{2}} - \frac{\epsilon}{2} \nabla_{\boldsymbol{x}} \mathcal{V}(\boldsymbol{x}_{t+1})$$

  This simple but precise interleaving of discrete momentum and position updates ensures exact volume preservation on phase space, and hence the accurate numerical trajectories we need to realize the potential of a Hamiltonian transition.

- Employing symplectic integrators provides the opportunity to translate the theoretical per-

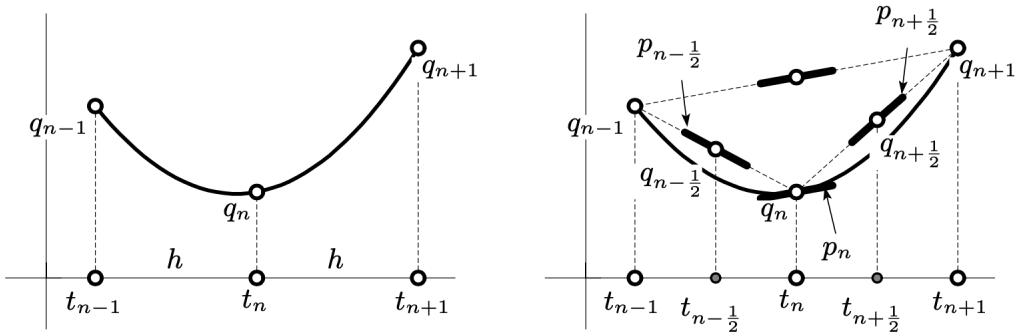

Fig. 1.6. Illustration for the Störmer–Verlet method

Figure 10: Leapfrog integrator can be seen as produced by parabolas [Haier et al., 2006]

formance of the Hamiltonian Monte Carlo method into a practical implementation. There remain, however, two obstructions to realizing this translation.

- First, even though symplectic integrators are highly accurate, the small errors they do introduce will bias the resulting Hamiltonian transitions without an exact correction.
- Second, we have to be able to select a symplectic integrator well-suited to a given target distribution.

3.2 Correcting for Symplectic Integrator Error

- One particularly natural strategy for correcting the bias introduced by the error in a symplectic integrator is to treat the ***Hamiltonian transition as the proposal*** for a Metropolis-Hastings scheme on phase space.
- Suppose $(\boldsymbol{x}_L, \boldsymbol{v}_L)$ be the last state of L symplectic integrator steps starting from $(\boldsymbol{x}_0, \boldsymbol{v}_0)$. We are interested in the *Hastings ratio* (i.e. the acceptance rate threshold):

$$r(\boldsymbol{x}_0, \boldsymbol{v}_0; \boldsymbol{x}_L, \boldsymbol{v}_L) = \frac{\pi(\boldsymbol{x}_L, \boldsymbol{v}_L)\, K((\boldsymbol{x}_L, \boldsymbol{v}_L); (\boldsymbol{x}_0, \boldsymbol{v}_0))}{\pi(\boldsymbol{x}_0, \boldsymbol{v}_0)\, K((\boldsymbol{x}_0, \boldsymbol{v}_0); (\boldsymbol{x}_L, \boldsymbol{v}_L))}$$

  where $K((\boldsymbol{x}_0, \boldsymbol{v}_0); (\boldsymbol{x}_L, \boldsymbol{v}_L)) := T\{(\boldsymbol{x}_0, \boldsymbol{v}_0) \to (\boldsymbol{x}_L, \boldsymbol{v}_L)\} = \delta_{\boldsymbol{x}_0}(\boldsymbol{x}_L)\, \delta_{\boldsymbol{v}_0}(\boldsymbol{v}_L)$ is the transition kernel (rule) from $(\boldsymbol{x}_0, \boldsymbol{v}_0)$ to $(\boldsymbol{x}_L, \boldsymbol{v}_L)$. For HMC, this transition is achieved via numerical integrator simulating the Hamilton dynamics.

  For a given integrator, because we can propose only states going forwards and not backwards, this ratio is always 0.

- If we modify the Hamiltonian transition to be ***reversible***, however, then the ratio of proposal densities becomes non-zero and we achieve a useful correction scheme. The simplest way of achieving a reversible proposal is to augment the the numerical integration with a negation step that ***flips the sign of momentum***. Thus the reverse proposal is

$$\begin{aligned} K((\boldsymbol{x}_L, -\boldsymbol{v}_L); (\boldsymbol{x}_0, \boldsymbol{v}_0)) &:= T\{(\boldsymbol{x}_L, -\boldsymbol{v}_L) \to (\boldsymbol{x}_0, \boldsymbol{v}_0)\} \\ &= \delta_{\boldsymbol{x}_L}(\boldsymbol{x}_0)\, \delta_{-\boldsymbol{v}_L}(\boldsymbol{v}_0) \end{aligned}$$

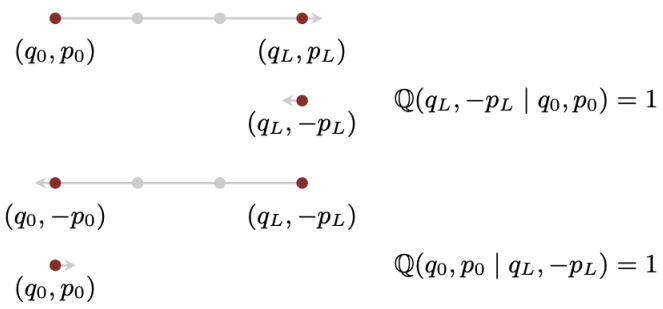

FIG 31. *Augmenting the numerical trajectory with a momentum flip defines a reversible Metropolis-Hastings proposal for which the forwards and backwards proposal probabilities are both well-behaved and we recover a valid correction scheme.*

Figure 11: Obtaining the reverse chain by flipping the sign of momentum. Note that the kinetic energy is symmetric to the change of sign in momentum. [Betancourt, 2017]

- The ***Hastings ratio*** is then computed as

$$\begin{aligned} r(\boldsymbol{x}_0, \boldsymbol{v}_0; \boldsymbol{x}_L, -\boldsymbol{v}_L) &= \frac{\pi(\boldsymbol{x}_L, -\boldsymbol{v}_L)\, K((\boldsymbol{x}_L, -\boldsymbol{v}_L); (\boldsymbol{x}_0, \boldsymbol{v}_0))}{\pi(\boldsymbol{x}_0, \boldsymbol{v}_0) K((\boldsymbol{x}_0, \boldsymbol{v}_0); (\boldsymbol{x}_L, -\boldsymbol{v}_L))} \\ &= \frac{\pi(\boldsymbol{x}_L, -\boldsymbol{v}_L)\, \delta_{\boldsymbol{x}_L}(\boldsymbol{x}_0)\, \delta_{-\boldsymbol{v}_L}(\boldsymbol{v}_0)}{\pi(\boldsymbol{x}_0, \boldsymbol{v}_0)\, \delta_{\boldsymbol{x}_0}(\boldsymbol{x}_L)\, \delta_{\boldsymbol{v}_0}(-\boldsymbol{v}_L)} \\ &= \frac{\pi(\boldsymbol{x}_L, -\boldsymbol{v}_L)}{\pi(\boldsymbol{x}_0, \boldsymbol{v}_0)} = \frac{\exp(-\mathcal{H}(\boldsymbol{x}_L, -\boldsymbol{v}_L))}{\exp(-\mathcal{H}(\boldsymbol{x}_0, \boldsymbol{v}_0))} \\ &= \exp\left(-\mathcal{H}(\boldsymbol{x}_L, -\boldsymbol{v}_L) + \mathcal{H}(\boldsymbol{x}_0, \boldsymbol{v}_0)\right) \end{aligned} \tag{27}$$

- Finally, we have ***Metropolized Hamiltonian Monte Carlo*** [Brooks et al., 2011, Betancourt, 2017]:

  1. For $t = 1, 2, \ldots, k$:

     (a) Generate a momentum $\boldsymbol{V}_{t-1}$ from Normal distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_v)$;

     (b) Set $(\boldsymbol{X}', \boldsymbol{V}') = \widehat{\phi}_T(\boldsymbol{X}_{t-1}, \boldsymbol{V}_{t-1})$ as the last state of T symplectic integrator steps starting from $(\boldsymbol{X}_{t-1}, \boldsymbol{V}_{t-1})$.

     (c) Compute the Hastings ratio:

$$r(\boldsymbol{X}_{t-1}, \boldsymbol{V}_{t-1}; \boldsymbol{X}', -\boldsymbol{V}') = \exp\left(-\mathcal{H}(\boldsymbol{X}', -\boldsymbol{V}') + \mathcal{H}(\boldsymbol{X}_{t-1}, \boldsymbol{V}_{t-1})\right)$$

     (d) Accept $\boldsymbol{X}_t = \boldsymbol{X}'$ with probability

$$\alpha(\boldsymbol{X}_{t-1}, \boldsymbol{V}_{t-1}; \boldsymbol{X}', -\boldsymbol{V}') = \min\left\{1,\ r(\boldsymbol{X}_{t-1}, \boldsymbol{V}_{t-1}; \boldsymbol{X}', -\boldsymbol{V}')\right\}$$

     (e) Otherwise, accept $\boldsymbol{X}_t = \boldsymbol{X}_{t-1}$.

- Symplectic integrators are not exactly energy preserving, causing their numerical trajectories to deviate from the target energy level set. In particular, sampling a state uniformly from any numerical trajectory will not generate a sample from the canonical distribution. This error, however, can be exactly corrected by sampling from the trajectory not uniformly but rather with weights proportional to the desired canonical density function [Betancourt, 2017].

- Regarding the ***robustness*** of the HMC, preliminary results show that even simple implementations of the Hamiltonian Monte Carlo method are geometrically ergodic over a large class of target distributions, larger than the class for non-gradient based algorithms like Random-walk Metropolis-Hastings.
- There also some concerns in practice:
  - In practice, ***longer trajectory*** T is preferred so that the end state is less correlated with the initial state. However, it will have larger cost of simulation time.
  - Also, if the ***kinetic energy*** is *poorly-chosen* then the marginal energy distribution can become heavy-tailed itself in which case the stochastic exploration between level sets will become so slow that after any finite number of transitions the exploration of the Markov chain will be incomplete.
  - Another common obstruction to geometric ergodicity is neighborhoods in parameter space where the target distribution exhibits ***large curvature***. Most Markov transitions are not able to resolve these narrow neighborhoods, resulting in incomplete exploration and biased Markov chain Monte Carlo estimators.
  - Finally, HMC requires computation on the ***gradients*** in the symplectic integrator. It would be computational expensive if the target distribution has complex form.

4 Connections with Information Geometry

- It can be established that the performance of HMC depends on the geometrical property of the target distributions.
- **Theorem 4.1** *[Vishnoi, 2021]*
  Let $\mathcal{V} : \mathbb{R}^d \to \mathbb{R}$ be a twice-differentiable function which satisfies $m\,\boldsymbol{I}_d \preceq \nabla^2_{\boldsymbol{x}}(\mathcal{V}(\boldsymbol{x})) \preceq M\,\boldsymbol{I}_d$. Let μ_k be the distribution of $\boldsymbol{X}_k$ at step k from ***the Idealized HMC****. Suppose that both μ_0 and π have mean and variance bounded by $\mathcal{O}(1)$. Then given any $\epsilon > 0$, for $T = \Omega(\frac{\sqrt{m}}{M})$ and $k = \mathcal{O}((M/m)^2 \log(1/\epsilon))$, we have that $\mathcal{W}_2(\mu_k, \pi) \leq \epsilon$.*

References

Michael Betancourt. A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*, 2017.

Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of markov chain monte carlo*. CRC press, 2011.

Izrail Moiseevitch Gelfand, Richard A Silverman, et al. *Calculus of variations*. Courier Corporation, 2000.

Ernst Haier, Christian Lubich, and Gerhard Wanner. *Geometric Numerical integration: structure-preserving algorithms for ordinary differential equations*. Springer, 2006.

Benedict Leimkuhler and Sebastian Reich. *Simulating hamiltonian dynamics*. Number 14. Cambridge university press, 2004.

Jun S Liu. *Monte Carlo strategies in scientific computing*, volume 10. Springer, 2001.

Nisheeth K Vishnoi. An introduction to hamiltonian monte carlo method for sampling. *arXiv preprint arXiv:2108.12107*, 2021.