Advanced Simulation Methods

Chapter 7 - Hamiltonian Monte Carlo

In this chapter, we are going to study an MCMC method called Hamiltonian Monte Carlo. This method is based on Hamiltonian dynamics, and allows efficient exploration of the state space by making large moves. It is known to be more efficient in high dimensions than Random Walk Metropolis-Hastings and MALA, due to the fact that it avoids random walk behaviour.

1 Hamiltonian mechanics

Let $\boldsymbol{q} \in \mathbb{R}^d$ denote the position, $\boldsymbol{p} \in \mathbb{R}^d$ denote the momentum, $U : \mathbb{R}^d \to \mathbb{R}$ be a continuously differentiable potential function, and $\boldsymbol{M} \in \mathbb{R}^{d\times d}$ be a symmetric positive definite matrix. Then Hamilton's equations are

$$\frac{d}{dt}\boldsymbol{q} = \boldsymbol{M}^{-1}\boldsymbol{p}, \tag{1}$$

$$\frac{d}{dt}\boldsymbol{p} = -\nabla U(\boldsymbol{q}). \tag{2}$$

In the case when $\boldsymbol{M} = m\boldsymbol{I}_d$, these equations correspond to Newtonian mechanics with a non-dissipative force arising from the potential field $U(\boldsymbol{q})$. Equations (1)-(2) are termed a Hamiltonian system, which has *Hamiltonian energy*

$$H(\boldsymbol{q},\boldsymbol{p}) := \frac{\boldsymbol{p}^T\boldsymbol{M}^{-1}\boldsymbol{p}}{2} + U(\boldsymbol{q}) = K(\boldsymbol{p}) + U(\boldsymbol{q}). \tag{3}$$

An important property of the Hamiltonian dynamics is that the Hamiltonian energy is preserved, since

$$\frac{d}{dt}H(\boldsymbol{q}(t),\boldsymbol{p}(t)) = \boldsymbol{p}(t)^T\boldsymbol{M}^{-1}\frac{d}{dt}\boldsymbol{p}(t) + \left(\frac{d}{dt}\boldsymbol{q}(t)\right)^T \nabla U(\boldsymbol{q}) = 0. \tag{4}$$

It is going to be very useful in our analysis to rewrite Hamiltonian equations in their canonical form

$$\frac{d}{dt}\boldsymbol{q} = +\nabla_{\boldsymbol{p}}H(\boldsymbol{q},\boldsymbol{p}), \tag{5}$$

$$\frac{d}{dt}\boldsymbol{p} = -\nabla_{\boldsymbol{q}}H(\boldsymbol{q},\boldsymbol{p}). \tag{6}$$

It is also going to be convenient to use the notation $\boldsymbol{z} := \begin{pmatrix}\boldsymbol{q}\\ \boldsymbol{p}\end{pmatrix}$, then $\boldsymbol{z} \in \mathbb{R}^{2d}$. Let $\boldsymbol{J}$ be the *canonical structure matrix*

$$\boldsymbol{J} := \begin{pmatrix}\boldsymbol{0} & \boldsymbol{I}_d\\ -\boldsymbol{I}_d & \boldsymbol{0}\end{pmatrix}, \tag{7}$$

then the canonical Hamiltonian equations (5)-(6) can be rewritten as

$$\frac{d}{dt}\boldsymbol{z} = \boldsymbol{J}\nabla_{\boldsymbol{z}}H(\boldsymbol{z}). \tag{8}$$

Example 1 (Harmonic oscillator) *Let* $\boldsymbol{L} = \begin{pmatrix}\omega^2 & 0\\ 0 & 1\end{pmatrix}$, *and* $H(\boldsymbol{z}) = \frac{1}{2}\boldsymbol{z}^T\boldsymbol{L}\boldsymbol{z} = \frac{1}{2}\omega^2q^2 + \frac{1}{2}p^2$, *and the Hamiltonian equations become*

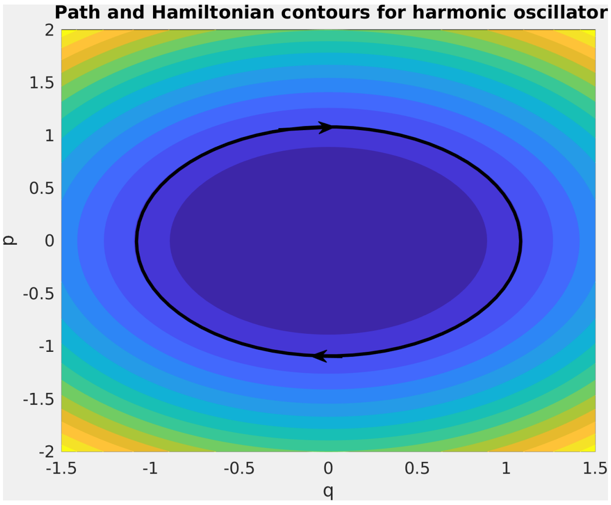
$$\frac{d}{dt}\boldsymbol{z} = \boldsymbol{J}\nabla_{\boldsymbol{z}}H(\boldsymbol{z}) = \boldsymbol{J}\boldsymbol{L}\boldsymbol{z}, \tag{9}$$

using the particular choice of $\boldsymbol{L}$, this is equivalent to

$$\frac{d}{dt}q = p \tag{10}$$
$$\frac{d}{dt}p = -\omega^2 q, \tag{11}$$

which corresponds to the harmonic oscillator. In general, (9) *can be defined for arbitrary* $2d \times 2d$ *symmetric matrix* $\boldsymbol{L}$*, and this is called a* linear Hamiltonian system*.*

The following figure illustrates the evolution of the Harmonic oscillator, and the preservation of the Hamiltonian energy $H(q,p) = \frac{1}{2}\omega^2 q^2 + \frac{1}{2}p^2$*, corresponding to an ellipse in the phase space.*



2 Symplecticity and other properties

A smooth map $\boldsymbol{\Psi} : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ is called *symplectic* if its Jacobian $\nabla_z \boldsymbol{\Psi}(z)$ satisfies that

$$[\nabla \boldsymbol{\Psi}(\boldsymbol{z})]^T \boldsymbol{J}^{-1} \nabla \boldsymbol{\Psi}(\boldsymbol{z}) = \boldsymbol{J}^{-1} \tag{12}$$

for all $\boldsymbol{z}$ in $\mathbb{R}^{2d}$, where $\boldsymbol{J} = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{I}_d \\ -\boldsymbol{I}_d & \boldsymbol{0} \end{pmatrix}$ is the canonical structure matrix. This satisfies that $\boldsymbol{J}^{-1} = \boldsymbol{J}^T$.

An important property of symplectic maps is the volume preservation property.

Proposition 1 *Symplectic maps are volume preserving.*

Proof. It can be shown that the infinitesimal cube $[z_1, z_1 + \delta] \times \ldots \times [z_{2d}, z_{2d} + \delta]$ will be mapped by $\boldsymbol{\Psi}$ into a $2d$ dimensional parallelepiped with volume $|\det(\nabla_{\boldsymbol{z}} \boldsymbol{\Psi}(\boldsymbol{z}))|\delta^{2d}$. Using the product rule for determinants on the symplectic condition (12), we have

$$\det(\nabla_z \boldsymbol{\Psi}(z))^2 \det(\boldsymbol{J}^{-1}) = \det(\boldsymbol{J}^{-1}),$$

which implies that $|\det(\nabla_z \boldsymbol{\Psi}(z))| = 1$ since $|\det(\boldsymbol{J}^{-1})| = 1$. Hence the infinitesimal volumes are preserved, and the same property can be obtained for non-infinitesimal volumes by integration. ■

A key property of Hamiltonian dynamics is that it is symplectic.

Proposition 2 *The Hamiltonian flow defined by* (8) *is symplectic and volume preserving.*

Proof. Let $\boldsymbol{\Psi}_{t,H}(\boldsymbol{z})$ denote the flow-map of the Hamiltonian dynamics (8) with Hamiltonian $H(\boldsymbol{z})$, i.e. $\boldsymbol{\Psi}_{t,H}(\boldsymbol{z}(0)) = \boldsymbol{z}(t)$ for a solution of (8). In order to show the symplecticity of the Hamiltonian dynamics, we need to understand the behaviour of the Jacobian $\frac{\partial}{\partial \boldsymbol{z}} \boldsymbol{\Psi}_{t,H}(\boldsymbol{z})$. Now we are going to describe the evolution of this Jacobian in time. Let $\boldsymbol{z}(0)$ be an initial point, and $\overline{\boldsymbol{z}}(0) = \boldsymbol{z}(0) + \delta \boldsymbol{z}(0)$ be another nearby initial point. Then the differences of two paths of the Hamiltonian dynamics at time t can be written as

$$\delta \boldsymbol{z}(t) := \overline{\boldsymbol{z}}(t) - \boldsymbol{z}(t) = \boldsymbol{\Psi}_{t,H}(\overline{\boldsymbol{z}}(0)) - \boldsymbol{\Psi}_{t,H}(\boldsymbol{z}(0)) = \frac{\partial}{\partial \boldsymbol{z}(0)} \boldsymbol{\Psi}_{t,H}(\boldsymbol{z}(0)) \cdot \delta \boldsymbol{z}(0) + o(\delta \boldsymbol{z}(0)),$$

and by differentiating the right hand side in t, we obtain that

$$\begin{aligned}\frac{d}{dt}\left[\frac{\partial}{\partial \boldsymbol{z}(0)}\boldsymbol{\Psi}_{t,H}(\boldsymbol{z}(0))\delta\boldsymbol{z}(0)\right] &= \frac{\partial}{\partial \boldsymbol{z}(0)}\left(\frac{d}{dt}\boldsymbol{\Psi}_{t,H}(\boldsymbol{z}(0))\right)\cdot\delta\boldsymbol{z}(0)\\ &= \frac{\partial}{\partial \boldsymbol{z}(0)}\left(\boldsymbol{J}\nabla H(\boldsymbol{z}(t))\right)\cdot\delta\boldsymbol{z}(0) = \frac{\partial}{\partial \boldsymbol{z}(0)}\left(\boldsymbol{J}\nabla H(\boldsymbol{\Psi}_{t,H}(\boldsymbol{z}(0)))\right)\cdot\delta\boldsymbol{z}(0)\\ &= \boldsymbol{J}\nabla^2 H(\boldsymbol{z}(t))\cdot\left[\frac{\partial}{\partial \boldsymbol{z}(0)}\boldsymbol{\Psi}_{t,H}(\boldsymbol{z}(0))\delta\boldsymbol{z}(0)\right].\end{aligned}$$

The above equation describing the evolution of $\frac{\partial}{\partial \boldsymbol{z}(0)}\boldsymbol{\Psi}_{t,H}(\boldsymbol{z}(0))\cdot\delta\boldsymbol{z}(0)$, the infinitesimal distance between two paths. It is called the *variational equation.* Let $\boldsymbol{F}(t) := \frac{\partial}{\partial \boldsymbol{z}(0)}\boldsymbol{\Psi}_{t,H}(\boldsymbol{z}(0))$ be the Jacobian matrix of the Hamiltonian flow at time t (for simplicity, the dependence on $\boldsymbol{z}(0)$ is supressed in the notation). Using the fact that the variational equations hold for every direction $\delta\boldsymbol{z}(0)$, it follows that this satisfies the matrix valued variational equation

$$\frac{d}{dt}\boldsymbol{F} = \boldsymbol{J}\nabla^2 H(\boldsymbol{z}(t))\boldsymbol{F}, \tag{13}$$

with initial condition $\boldsymbol{F}(0) = \boldsymbol{I}_{2d}$.

In order to establish that Hamiltonian flow is symplectic, it suffices to verify that (12) holds. In this case this is equivalent to

$$\boldsymbol{F}_t^T\boldsymbol{J}^{-1}\boldsymbol{F}_t = \boldsymbol{J}^{-1}. \tag{14}$$

Since $\boldsymbol{F}_0 = \boldsymbol{I}_{2d}$, (14) holds for $t = 0$. By taking the derivative, and using (13) we obtain that

$$\begin{aligned}\frac{d}{dt}(\boldsymbol{F}_t^T\boldsymbol{J}^{-1}\boldsymbol{F}_t) &= (\boldsymbol{J}\nabla^2 H(\boldsymbol{z}(t))\boldsymbol{F}_t)^T\boldsymbol{J}^{-1}\boldsymbol{F}_t + \boldsymbol{F}_t\boldsymbol{J}^{-1}\boldsymbol{J}\nabla^2 H(\boldsymbol{z}(t))\boldsymbol{F}_t\\ &= \boldsymbol{F}_t^T\nabla^2 H(\boldsymbol{z}(t))\boldsymbol{J}^T\boldsymbol{J}^{-1}\boldsymbol{F}_t + \boldsymbol{F}_t^T\boldsymbol{J}^{-1}\boldsymbol{J}\nabla^2 H(\boldsymbol{z}(t))\boldsymbol{F}_t = \boldsymbol{0},\end{aligned}$$

since $\boldsymbol{J}^T\boldsymbol{J}^{-1} = -\boldsymbol{J}\boldsymbol{J}^{-1} = -\boldsymbol{I}_{2d}$ and $\boldsymbol{J}^{-1}\boldsymbol{J} = \boldsymbol{I}_{2d}$. Hence (14) holds for every $t \geq 0$ and the Hamiltonian flow is symplectic. Since every symplectic flow is volume preserving, the Hamiltonian flow is also volume preserving. ■

Finally, the next proposition states the stationary of the target with respect to Hamiltonian flow.

Proposition 3 *The distribution $\pi(d\boldsymbol{z}) \propto \exp(-H(\boldsymbol{z}))d\boldsymbol{z}$ is stationary with respect to the Hamiltonian flow.*

Proof. Let the density of π be equal $\pi(\boldsymbol{z}) = \frac{\exp(-H(\boldsymbol{z}))}{C}$ for some normalising constant C. To see that this result holds, let $\boldsymbol{Z}_0 \sim \pi$, and $\boldsymbol{Z}_t = \boldsymbol{\Psi}_{t,H}(\boldsymbol{Z}(0))$. Let $\boldsymbol{z} = (z_1, \ldots, z_d)^T$, then for infinitesimally small δ,

$$\begin{aligned}&\mathbb{P}(\boldsymbol{Z}_t \in [z_1, z_1+\delta] \times \ldots \times [z_d, z_d+\delta]) = \mathbb{P}\left(\boldsymbol{Z}_0 \in \boldsymbol{\Psi}_{-t,H}\left([z_1, z_1+\delta] \times \ldots \times [z_d, z_d+\delta]\right)\right)\\ &= \pi(\boldsymbol{\Psi}_{-t,H}(\boldsymbol{z}))\mathrm{Vol}\left(\boldsymbol{\Psi}_{-t,H}\left([z_1, z_1+\delta] \times \ldots \times [z_d, z_d+\delta]\right)\right)\\ &= \frac{\exp(-H(\boldsymbol{\Psi}_{-t,H}(\boldsymbol{z})))}{C}\mathrm{Vol}\left(\boldsymbol{\Psi}_{-t,H}\left([z_1, z_1+\delta] \times \ldots \times [z_d, z_d+\delta]\right)\right)\end{aligned}$$

using the preservation of the Hamiltonian, and the volume,

$$= \frac{\exp(-H(\boldsymbol{z}))}{C}\mathrm{Vol}\left([z_1, z_1+\delta] \times \ldots \times [z_d, z_d+\delta]\right) = \pi([z_1, z_1+\delta] \times \ldots \times [z_d, z_d+\delta]),$$

hence the stationarity is established. ■

Hamilton's equations themselves do not define an ergodic Markov chain, as they keep the Hamiltonian invariant. Let $\boldsymbol{P}_R$ be the Markov kernel that resample the momentum according to the Gaussian distribution with covariance matrix $\boldsymbol{M}$, and $\boldsymbol{\Psi}_{T,H}$ be the Hamiltonian flow-map for time T. Then π is stationary with respect to the combination $\boldsymbol{P}_R\cdot\boldsymbol{\Psi}_{T,H}$ (i.e. first resample momentum, then move according to the Hamiltonian dynamics). Moreover, it is easy to show that if we choose T to be a random variable chosen independently at teach time step, then the resulting Markov kernel is a mixture of kernels whose stationary distribution is π, and therefore it also admits π as a stationary distribution. We call this Markov kernel *Randomized Hamiltonian Monte Carlo* with continuous dynamics. The following proposition states some conditions that guarantee π-irreducibility.

Proposition 4 *Let T be distributed according to ν_T, which has positive density on an interval $[0,\tau]$ for some $\tau > 0$. Let $\boldsymbol{K}$ denote the Markov kernel for the position variables on $\mathbb{R}^d$ corresponding to sampling a random momentum $\boldsymbol{p}$, then running the Hamiltonian dynamics started at $(\boldsymbol{q},\boldsymbol{p})$ up to time T sampled from ν_T (independently at each step), and finally discarding the momentum variable. Suppose that U is continuously differentiable on $\mathbb{R}^d$, and satisfies that $\sup_{\boldsymbol{q}} \|\nabla^2 U(\boldsymbol{q})\| \le L$, and $\inf_{\boldsymbol{q}\in\mathbb{R}^d} U(\boldsymbol{q}) > -\infty$. Then $\boldsymbol{K}$ is strongly ν-irreducible.*

In order to show this result, we are going to first show two preliminary lemmas.

Lemma 1 *In addition to the conditions of Proposition 4, suppose that $\boldsymbol{M} = \boldsymbol{I}_d$, and that $0 \le t \le \frac{1}{2(L+1)^3}$. For any starting point $\boldsymbol{q}(0), \boldsymbol{p}(0)$, let $(\boldsymbol{q}(s), \boldsymbol{p}(s))_{s\ge 0}$ be the solution of the Hamiltonian dynamics (1)-(2). Then the Jacobian of $\boldsymbol{q}(t)$ in terms of the initial momentum $\boldsymbol{p}(0)$ satisfies that*

$$\frac{1}{2} t \boldsymbol{I}_d \preceq \frac{\partial \boldsymbol{q}(t)}{\partial \boldsymbol{p}(0)} \preceq \frac{3}{2} t \boldsymbol{I}_d,$$

where $\preceq$ denotes the positive semidefinite order (i.e. $\boldsymbol{A} \preceq \boldsymbol{B}$ means that $\boldsymbol{B} - \boldsymbol{A}$ is positive semidefinite).

Proof. Let $\boldsymbol{F}(t) := \frac{\partial}{\partial \boldsymbol{z}(0)} \boldsymbol{\Psi}_{t,H}(\boldsymbol{z}(0))$ denote the Jacobian matrix at time t, then from the evolution equation (13) it follows that

$$\begin{aligned}\frac{d}{dt}(\boldsymbol{F} - \boldsymbol{I}_{2d}) &= \boldsymbol{J}\nabla^2 H(\boldsymbol{z}(t))\boldsymbol{F} = \boldsymbol{J}\nabla^2 H(\boldsymbol{z}(t))(\boldsymbol{F} - \boldsymbol{I}_{2d}) + \boldsymbol{J}\nabla^2 H(\boldsymbol{z}(t)) \\ &= \begin{pmatrix} \boldsymbol{0} & \boldsymbol{I}_d \\ -\nabla^2 U(\boldsymbol{q}_t) & \boldsymbol{0} \end{pmatrix} ((\boldsymbol{F}(t) - \boldsymbol{I}_{2d}) + \boldsymbol{I}_{2d}). \qquad (15)\end{aligned}$$

Using the assumption of the lemma we know that $\|\nabla^2 U(\boldsymbol{q}_t)\| \le L$, hence $\left\|\begin{pmatrix} \boldsymbol{0} & \boldsymbol{I}_d \\ -\nabla^2 U(\boldsymbol{q}_t) & \boldsymbol{0} \end{pmatrix}\right\| \le (L+1)$, and therefore by (15) we have $\frac{d}{dt}(\|\boldsymbol{F} - \boldsymbol{I}_{2d}\| + (L+1)) \le (L+1)(\|\boldsymbol{F} - \boldsymbol{I}_{2d}\| + (L+1))$. Using Gronwall's lemma, and the fact that $\boldsymbol{F}(0) = \boldsymbol{I}_d$, we have $\|\boldsymbol{F}(t) - \boldsymbol{I}_{2d}\| + (L+1) \le (L+1)\exp((L+1)t)$. Hence for $t \le \frac{1}{L+1}$, we have

$$\|\boldsymbol{F}(t) - \boldsymbol{I}_{2d}\| \le (L+1)(\exp((L+1)t) - 1) \le 2(L+1)^2 t.$$

We can rewrite (15) as

$$\frac{d}{dt}\left(\boldsymbol{F} - \boldsymbol{I}_{2d} - \begin{pmatrix} \boldsymbol{0} & t\boldsymbol{I}_d \\ -\int_{s=0}^t \nabla^2 U(\boldsymbol{q}_s) ds & \boldsymbol{0} \end{pmatrix}\right) = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{I}_d \\ -\nabla^2 U(\boldsymbol{q}_t) & \boldsymbol{0} \end{pmatrix} (\boldsymbol{F}(t) - \boldsymbol{I}_{2d}),$$

where $\left\|\begin{pmatrix} \boldsymbol{0} & \boldsymbol{I}_d \\ -\nabla^2 U(\boldsymbol{q}_t) & \boldsymbol{0} \end{pmatrix} (\boldsymbol{F}(t) - \boldsymbol{I}_{2d})\right\| \le 2(L+1)^3 t$ for $t \le \frac{1}{L+1}$. Since $\frac{\partial \boldsymbol{q}(t)}{\partial \boldsymbol{p}(0)}$ is the upper-right block matrix of $\boldsymbol{F}(t)$, the above equation implies that $\|\frac{\partial \boldsymbol{q}(t)}{\partial \boldsymbol{p}(0)} - t\boldsymbol{I}_d\| \le \int_{s=0}^t s \cdot 2(L+1)^3 ds = (L+1)^3 t^2$, and the result follows. ■

Lemma 2 *Under the conditions of Proposition 4, and assuming that $\boldsymbol{M} = \boldsymbol{I}_d$, for any $\boldsymbol{q}(0), \boldsymbol{Q} \in \mathbb{R}^d$, any $0 < t \le \tau$, if $\boldsymbol{p}(0) = \frac{\boldsymbol{Q} - \boldsymbol{q}(0)}{t}$, and $(\boldsymbol{q}(s), \boldsymbol{p}(s))_{s\ge0}$ is the solution of the Hamiltonian dynamics (1)-(2), then*

$$\|\boldsymbol{q}(t) - \boldsymbol{Q}\| \le Ct^2$$

for a finite constant C depending only on $\boldsymbol{q}$, $\boldsymbol{Q}$, τ and U but independent of t.

Proof. Let $H(\boldsymbol{q}(0), \boldsymbol{p}(0)) = U(\boldsymbol{q}(0)) + \frac{\|\boldsymbol{p}(0)\|^2}{2}$ denote the initial Hamiltonian, and $H(\boldsymbol{q}(s), \boldsymbol{p}(s)) = U(\boldsymbol{q}(s)) + \frac{\|\boldsymbol{p}(s)\|^2}{2}$ be the current Hamiltonian. Since this is preserved by the Hamiltonian dynamics, for any $s \ge 0$, we have

$$\frac{\|\boldsymbol{p}(s)\|^2}{2} = U(\boldsymbol{q}) + \frac{\|\boldsymbol{p}(0)\|^2}{2} - U(\boldsymbol{q}(s)) \le \frac{\|\boldsymbol{p}(0)\|^2}{2} + U(\boldsymbol{q}) - \inf_{\boldsymbol{q}'\in\mathbb{R}^d} U(\boldsymbol{q}'),$$

$$\|\boldsymbol{p}(s)\| \le \sqrt{2(U(\boldsymbol{q}) - \inf_{\boldsymbol{q}'\in\mathbb{R}^d} U(\boldsymbol{q}')) + \|\boldsymbol{p}(0)\|^2} \le \sqrt{2(U(\boldsymbol{q}(0)) - \inf_{\boldsymbol{q}'\in\mathbb{R}^d} U(\boldsymbol{q}'))} + \|\boldsymbol{p}(0)\|.$$

Since $\frac{d\boldsymbol{q}}{ds} = \boldsymbol{p}(s)$, we have

$$\|\boldsymbol{q}(s) - \boldsymbol{q}(0)\| \le \int_{r=0}^{s} \|\boldsymbol{p}(r)\| \mathrm{d}r \le s \cdot \left(\sqrt{2(U(\boldsymbol{q}(0)) - \inf_{\boldsymbol{q}' \in \mathbb{R}^d} U(\boldsymbol{q}'))} + \|\boldsymbol{p}(0)\| \right).$$

Hence for any $0 < t \le \tau$, for $\boldsymbol{p}(0) = \frac{\boldsymbol{Q} - \boldsymbol{q}(0)}{t}$, for any $0 \le s \le t$, we have

$$\|\boldsymbol{q}(s) - \boldsymbol{q}(0)\| \le D \text{ for } D := \tau \sqrt{2(U(\boldsymbol{q}) - \inf_{\boldsymbol{q}' \in \mathbb{R}^d} U(\boldsymbol{q}'))} + \|\boldsymbol{Q} - \boldsymbol{q}(0)\|. \tag{16}$$

Let $E := \sup_{\boldsymbol{q}: \|\boldsymbol{q} - \boldsymbol{q}(0)\| \le D} \|\nabla U(\boldsymbol{q})\|$, then using the assumption that $\sup_{\boldsymbol{q} \in \mathbb{R}^d} \|\nabla^2 U(\boldsymbol{q})\| \le L$, it follows that E is finite. Since by the Hamiltonian equations $\frac{d\boldsymbol{p}}{ds} = -\nabla U(\boldsymbol{q}(s))$, it follows that for $0 \le s \le t \le \tau$, we have $\|\frac{d\boldsymbol{p}}{ds}\| \le E$, and therefore

$$\|\boldsymbol{p}(s) - \boldsymbol{p}(0)\| = \left\| \int_{r=0}^{s} \frac{d\boldsymbol{p}}{ds} ds \right\| \le \int_{r=0}^{s} \left\| \frac{d\boldsymbol{p}}{ds} \right\| ds \le s \cdot E.$$

Now using the Hamiltonian equations again in q, we have that for $0 \le t \le \tau$,

$$\begin{aligned} \boldsymbol{q}(t) &= \boldsymbol{q}(0) + \int_{s=0}^{t} \boldsymbol{p}(s) ds = \boldsymbol{q}(0) + \int_{s=0}^{t} [p(0) + (p(s) - p(0))] ds \\ &= \boldsymbol{q}(0) + \boldsymbol{p}(0)t + \int_{s=0}^{t} (\boldsymbol{p}(s) - \boldsymbol{p}(0)) ds \end{aligned}$$

using the fact that $\boldsymbol{p}(0) = \frac{\boldsymbol{Q} - \boldsymbol{q}(0)}{t}$

$$= \boldsymbol{Q} + \int_{s=0}^{t} (\boldsymbol{p}(s) - \boldsymbol{p}(0)) ds.$$

Therefore we have

$$\|\boldsymbol{q}(t) - \boldsymbol{Q}\| = \left\| \int_{s=0}^{t} (\boldsymbol{p}(s) - \boldsymbol{p}(0)) ds \right\| \le \int_{s=0}^{t} \|\boldsymbol{p}(s) - \boldsymbol{p}(0)\| ds \le \int_{s=0}^{t} s \cdot E \mathrm{d}s = \frac{E}{2} t^2,$$

hence the claim of the lemma holds for $C = E/2$. ■

Proof of Proposition 4. Let $\nu(\boldsymbol{q}) \propto \exp(-U(\boldsymbol{q}))$ be the target distribution, and μ be the multivariate normal distribution on $\mathbb{R}^d$ with mean 0 and covariance matrix $\boldsymbol{M}$. Assume without loss of generality that $\boldsymbol{M} = \boldsymbol{I}_d$ (the general case follows by a simple modification of this argument).

Our goal is to show that for any starting point $\boldsymbol{q}(0)$, any measurable set $A \subset \mathbb{R}^d$ with $\nu(A) > 0$, we have $K(\boldsymbol{q}(0), A) > 0$. Since we can fill $\mathbb{R}^d$ into countably many balls of radius δ (for any $\delta > 0$), we can assume without loss of generality that A is a subset of a ball of radius δ centered at some point $\boldsymbol{Q} \in \mathbb{R}^d$. We will denote this closed ball by $B_\delta(\boldsymbol{Q}) = \{\boldsymbol{q} \in \mathbb{R}^d : \|\boldsymbol{q} - \boldsymbol{Q}\| \le \delta\}$.

Then by Lemma 2, it follows that if $t \le \frac{\sqrt{\delta}}{\sqrt{C}}$, and

$$\boldsymbol{p}(0) = \boldsymbol{p}^* = \frac{\boldsymbol{Q} - \boldsymbol{q}(0)}{t},$$

then $\boldsymbol{q}(t) \in B_\delta(\boldsymbol{Q})$.

Let $\boldsymbol{\Psi}(\boldsymbol{p}(0))$ denote the position ($\boldsymbol{q}$) component of $\boldsymbol{\Psi}_t(\boldsymbol{q}(0), \boldsymbol{p}(0))$ ($\boldsymbol{q}(0)$ is assumed fixed). $\boldsymbol{\Psi}$ maps the initial momentum $\boldsymbol{p}(0)$ to the position at time t.

It follows from Lemma 1 that for any vector $\boldsymbol{v}$, we have

$$\|\boldsymbol{\Psi}(\boldsymbol{p}^*) - \boldsymbol{\Psi}(\boldsymbol{p}^* + \boldsymbol{v})\| = \left\| \left(\int_{s=0}^{1} \frac{\partial \boldsymbol{q}(t)}{\partial \boldsymbol{p}(0)} \bigg|_{\boldsymbol{p}(0) = \boldsymbol{p}^* + \boldsymbol{v}s} ds \right) \boldsymbol{v} \right\| \ge \frac{t\|\boldsymbol{v}\|}{2}.$$

Therefore, with the choice $r = \frac{6\delta}{t}$ the sphere $S_r(\boldsymbol{p}^*) = \{\boldsymbol{p}' \in \mathbb{R}^d : \|\boldsymbol{p}' - \boldsymbol{p}^*\| = r\}$ satisfies that for any $\boldsymbol{p}' \in S_r(\boldsymbol{p}^*)$,

$$\|\boldsymbol{\Psi}(\boldsymbol{p}^*) - \boldsymbol{\Psi}(\boldsymbol{p})\| \ge 3\delta.$$

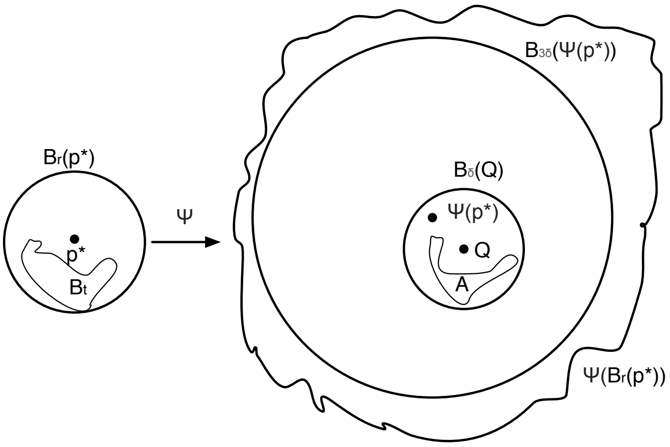


Figure 1: Effect of the map $\boldsymbol{\Psi}$

Since $\boldsymbol{\Psi}(\boldsymbol{p}^*)$ is contained in $B_\delta(\boldsymbol{Q})$, it follows that the map of the ball $B_r(\boldsymbol{p}(0))$ by $\boldsymbol{\Psi}$ will contain the ball $B_\delta(\boldsymbol{Q})$ (see Figure 1).

Hence for any point in $\boldsymbol{Q}' \in B_\delta(\boldsymbol{Q})$, there is at least one point $\boldsymbol{p}' \in B_r(\boldsymbol{p}(0))$ such that $\boldsymbol{\Psi}(\boldsymbol{p}') = \boldsymbol{q}'$. Denote the set of points in $B_r(\boldsymbol{p}(0))$ that get mapped into the set A by $\boldsymbol{\Psi}$ as B_t (preimage of A). Since $\nu(A) > 0$, and ν has a density, it follows that $\mathrm{Vol}(A) > 0$, and by Lemma 1, it follows that if $0 < t \le \frac{1}{2(L+1)^3}$, the determinant of the Jacobian of $\boldsymbol{\Psi}$ on B_t is finite and positive, so we also must have $\mathrm{Vol}(B_t) > 0$, implying that $\mu(B_t) > 0$ (μ denotes the standard normal distribution on $\mathbb{R}^d$, corresponding to the distribution of the resampled momentum variable). Since this holds for any $t \le \min\left(\frac{\sqrt{\delta}}{\sqrt{C}}, \frac{1}{2(L+1)^3}\right)$, and $T \sim \nu_T$ has positive density on $[0,\tau]$, it follows by integration that $K(\boldsymbol{q}(0), A) > 0$, thus K is strongly ν-irreducible. ■

In general the Hamiltonian flow cannot be simulated exactly due to the nonlinearity of the ODE. Hence we are going to approximate it by discretization.

3 Discretizing Hamilton's equations

We consider 3 simple discretization schemes.
Explicit scheme:

$$\begin{aligned}\boldsymbol{p}(t+\epsilon) &= \boldsymbol{p}(t) - \epsilon\nabla U(\boldsymbol{q}(t))\\ \boldsymbol{q}(t+\epsilon) &= \boldsymbol{q}(t) + \epsilon\boldsymbol{M}^{-1}\boldsymbol{p}(t).\end{aligned}$$

Modified explicit scheme:

$$\begin{aligned}\boldsymbol{p}(t+\epsilon) &= \boldsymbol{p}(t) - \epsilon\nabla U(\boldsymbol{q}(t))\\ \boldsymbol{q}(t+\epsilon) &= \boldsymbol{q}(t) + \epsilon\boldsymbol{M}^{-1}\boldsymbol{p}(t+\epsilon).\end{aligned}$$

Leapfrog (Störmer-Verlet) scheme

$$\begin{aligned}\boldsymbol{p}(t+\epsilon/2) &= \boldsymbol{p}(t) - \frac{\epsilon}{2}\nabla U(\boldsymbol{q}(t))\\ \boldsymbol{q}(t+\epsilon) &= \boldsymbol{q}(t) + \epsilon\boldsymbol{M}^{-1}\boldsymbol{p}(t+\epsilon/2).\\ \boldsymbol{p}(t+\epsilon) &= \boldsymbol{p}(t+\epsilon/2) - \frac{\epsilon}{2}\nabla U(\boldsymbol{q}(t+\epsilon))\end{aligned}$$

The following figure shows these schemes implemented for the 1 dimensional standard Gaussian distribution corresponding to $H(p,q) = \frac{p^2}{2} + \frac{q^2}{2}$.

As Figure 2 shows, these 3 schemes behave very differently. The first explicit scheme diverges for step size $\epsilon = 0.3$. The modified explicit scheme does not diverge, but it also does not tract the true trajectory closely. The Leapfrog scheme tracks the true trajectory very closely, and it does not diverge.

An important property of these discretizations is that they are symplectic, and hence volume preserving.

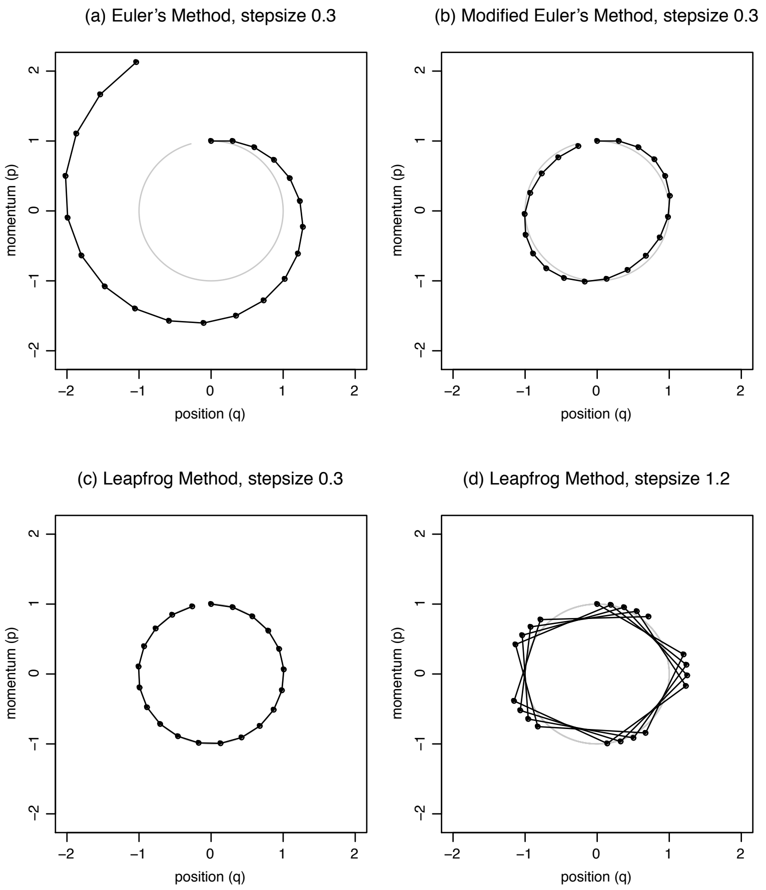


Figure 2: Approximation of Hamiltonian dynamics for $H(p,q) = \frac{p^2}{2} + \frac{q^2}{2}$ by 3 schemes. 20 steps in each case in black, the true trajectory in grey.

Proposition 5 *The above 3 schemes are symplectic, and hence volume preserving.*

Proof. First, note that by the chain rule for Jacobians, if $\boldsymbol{\Psi}(\boldsymbol{z}) = \boldsymbol{\Psi}_1(\boldsymbol{\Psi}_2(\boldsymbol{z}))$, then $\nabla_{\boldsymbol{z}}\boldsymbol{\Psi}(\boldsymbol{z}) = \nabla\boldsymbol{\Psi}_1(\boldsymbol{\Psi}_2(\boldsymbol{z})))\cdot \nabla\boldsymbol{\Psi}_2(\boldsymbol{z}))$. If $\boldsymbol{\Psi}_1$ and $\boldsymbol{\Psi}_2$ are symplectic, then

$$[\nabla\boldsymbol{\Psi}(\boldsymbol{z})]^T\boldsymbol{J}^{-1}\nabla_{\boldsymbol{z}}\boldsymbol{\Psi}(\boldsymbol{z}) = (\nabla\boldsymbol{\Psi}_2(\boldsymbol{z}))^T\left(\nabla\boldsymbol{\Psi}_1(\boldsymbol{\Psi}_2(\boldsymbol{z}))\right)^T\boldsymbol{J}^{-1}\nabla\boldsymbol{\Psi}_1(\boldsymbol{\Psi}_2(\boldsymbol{z}))\cdot\nabla\boldsymbol{\Psi}_2(\boldsymbol{z})) = \boldsymbol{J}^{-1},$$

so $\boldsymbol{\Psi}$ is also symplectic. By induction, this also holds for the composition of more than two maps. Using this fact, it suffices to check symplecticity for a single step in the above schemes, modifying only one of p or q. For example, the step $\boldsymbol{p}(t+\epsilon) = \boldsymbol{p}(t) - \epsilon\nabla U(\boldsymbol{q}(t))$ corresponds to the map $\boldsymbol{\Psi}(\boldsymbol{z}) = \begin{pmatrix}\boldsymbol{q}\\ \boldsymbol{p} - \epsilon\nabla U(\boldsymbol{q})\end{pmatrix}$, which has Jacobian $\nabla\boldsymbol{\Psi}(\boldsymbol{z}) = \begin{pmatrix}\boldsymbol{I}_d & \boldsymbol{0}\\ -\epsilon\nabla^2U(\boldsymbol{q}) & \boldsymbol{I}_d\end{pmatrix}$. Now by direct calculation, one can see that

$$(\nabla\boldsymbol{\Psi}(\boldsymbol{z}))^T\boldsymbol{J}^{-1}\nabla\boldsymbol{\Psi}(\boldsymbol{z}) = \begin{pmatrix}\boldsymbol{I}_d & -\epsilon\nabla^2U(\boldsymbol{q})\\ \boldsymbol{0} & \boldsymbol{I}_d\end{pmatrix}\begin{pmatrix}\boldsymbol{0} & -\boldsymbol{I}_d\\ \boldsymbol{I}_d & \boldsymbol{0}\end{pmatrix}\begin{pmatrix}\boldsymbol{I}_d & \boldsymbol{0}\\ -\epsilon\nabla^2U(\boldsymbol{q}) & \boldsymbol{I}_d\end{pmatrix} = \begin{pmatrix}\boldsymbol{0} & -\boldsymbol{I}_d\\ \boldsymbol{I}_d & \boldsymbol{0}\end{pmatrix} = \boldsymbol{J}^{-1},$$

hence the symplecticity holds. This can be also checked for the other steps, and hence all 3 schemes are symplectic. The volume preservation follows by Proposition 1. ■

4 Hamiltonian Monte Carlo

In this section, we are going to introduce Hamiltonian Monte Carlo, originally proposed in [1]. This is an MCMC method that samples from a target distribution with density $\mu(\boldsymbol{q}) \propto \exp(-U(\boldsymbol{q}))$ on $\mathbb{R}^d$ by extending the state space and sampling from the distribution π on $\mathbb{R}^{2d}$ with density $\pi(\boldsymbol{z}) \propto \exp(-H(\boldsymbol{z})) =$

$\exp(-U(\boldsymbol{q})-K(\boldsymbol{p}))=\exp(-U(\boldsymbol{q}))\exp\left(-\frac{\boldsymbol{p}^T\boldsymbol{M}^{-1}\boldsymbol{p}}{2}\right)$. From this decomposition, it is clear that the position and momentum variables are independent.

Let $\boldsymbol{\Psi}_\epsilon:\mathbb{R}^{2d}\to\mathbb{R}^{2d}$ denote the Leapfrog map, and for some $L\in\mathbb{N}$, we denote by $\boldsymbol{\Psi}_\epsilon^L$ the L times composition of the Leapfrog map. Let $\boldsymbol{N}:\mathbb{R}^{2d}\to\mathbb{R}^{2d}$ denote the map that negates the momentum, i.e. $\boldsymbol{N}\begin{pmatrix}\boldsymbol{q}\\ \boldsymbol{p}\end{pmatrix}=\begin{pmatrix}\boldsymbol{q}\\ -\boldsymbol{p}\end{pmatrix}$, and finally $\boldsymbol{P}_R$ be the Markov kernel that resamples the momentum component $\boldsymbol{p}$ from a multivariate Gaussian distribution with covariance matrix $\boldsymbol{M}$. Hamiltonian Monte Carlo consists of the iteration of 2 steps,

1. resample the momentum component $\boldsymbol{p}$, i.e. apply the Markov kernel $\boldsymbol{P}_R$.

2. We propose a new position $\begin{pmatrix}\boldsymbol{q}^*\\ \boldsymbol{p}^*\end{pmatrix}=\boldsymbol{\Psi}\begin{pmatrix}\boldsymbol{q}\\ \boldsymbol{p}\end{pmatrix}:=\boldsymbol{N}\left(\boldsymbol{\Psi}_\epsilon^L\begin{pmatrix}\boldsymbol{q}\\ \boldsymbol{p}\end{pmatrix}\right)$ by applying L Leapfrog steps and then flipping the momentum. This new position is accepted with probability

$$\min\left[1,\exp(H(\boldsymbol{q},\boldsymbol{p})-H(\boldsymbol{q}^*,p^*))\right]=\min\left[1,\exp\left(U(\boldsymbol{q})-U(\boldsymbol{q}^*)+\frac{1}{2}\boldsymbol{p}^T\boldsymbol{M}^{-1}\boldsymbol{p}-\frac{1}{2}(\boldsymbol{p}^*)^T\boldsymbol{M}^{-1}\boldsymbol{p}^*\right)\right]. \tag{17}$$

The following proposition shows the invariance of the target distribution for HMC.

Proposition 6 *π is invariant with respect to the Markov kernel proposed above. Moreover, π is reversible with respect to the second step of the Markov kernel.*

Proof. Since $\boldsymbol{q}$ and $\boldsymbol{p}$ are independent according to the target distribution π, it is clear that the first step keeps the target invariant. Let $\boldsymbol{P}_2$ denote the Markov kernel corresponding to the second step (a combination of a deterministic step and Metropolis-Hastings accept-reject step), then we are going to check that π is reversible with respect to $\boldsymbol{P}_2$.

We have seen during the previous lectures that in general a Markov kernel K is reversible with respect to a distribution π on state space $\mathbb{X}$ if for every bounded measurable function $f:\mathbb{X}^2\to\mathbb{R}$, we have

$$\int\int f(\boldsymbol{x},\boldsymbol{y})\pi(d\boldsymbol{x})K(\boldsymbol{x},d\boldsymbol{y})=\int\int f(\boldsymbol{x},\boldsymbol{y})\pi(d\boldsymbol{y})K(\boldsymbol{y},d\boldsymbol{x}). \tag{18}$$

In our case, $\boldsymbol{P}_2(\boldsymbol{x},d\boldsymbol{y})$ is non-zero only for $\boldsymbol{y}=\boldsymbol{\Psi}(\boldsymbol{x})$ and $\boldsymbol{y}=\boldsymbol{x}$, so we have

$$\begin{aligned}\int\int f(\boldsymbol{x},\boldsymbol{y})\pi(d\boldsymbol{x})\boldsymbol{P}_2(\boldsymbol{x},d\boldsymbol{y})=&\int\int f(\boldsymbol{x},\boldsymbol{\Psi}(\boldsymbol{x}))\min[1,e^{H(\boldsymbol{x})-H(\boldsymbol{\Psi}(\boldsymbol{x}))}]\pi(d\boldsymbol{x})\\ &+\int\int f(\boldsymbol{x},\boldsymbol{x})\left(1-\min[1,e^{H(\boldsymbol{x})-H(\boldsymbol{\Psi}(\boldsymbol{x}))}]\right)\pi(d\boldsymbol{x})\end{aligned}$$

Let $\boldsymbol{y}=\boldsymbol{\Psi}(\boldsymbol{x})$, then $\boldsymbol{x}=\boldsymbol{\Psi}(\boldsymbol{y})=\boldsymbol{\Psi}(\boldsymbol{\Psi}(\boldsymbol{x}))$, and by the volume preserving property of $\boldsymbol{\Psi}$, we have

$$\pi(d\boldsymbol{y})=\pi(d\boldsymbol{x})\cdot\frac{\exp(-H(\boldsymbol{y}))}{\exp(-H(\boldsymbol{x}))}=\pi(d\boldsymbol{x})\cdot e^{H(\boldsymbol{x})-H(\boldsymbol{y})},$$

and the first part of the above sum can be written

$$\begin{aligned}&\int\int f(\boldsymbol{x},\boldsymbol{\Psi}(\boldsymbol{x}))\min[1,e^{H(\boldsymbol{x})-H(\boldsymbol{\Psi}(\boldsymbol{x}))}]\pi(d\boldsymbol{x})\\ &=\int\int f(\boldsymbol{\Psi}(\boldsymbol{y}),\boldsymbol{y})\min[1,e^{H(\boldsymbol{\Psi}(\boldsymbol{y}))-H(\boldsymbol{y})}]\pi(d\boldsymbol{x})\\ &=\int\int f(\boldsymbol{\Psi}(\boldsymbol{y}),\boldsymbol{y})\min[1,e^{H(\boldsymbol{\Psi}(\boldsymbol{y}))-H(\boldsymbol{y})}]\cdot e^{H(\boldsymbol{y})-H(\boldsymbol{\Psi}(\boldsymbol{y}))}\pi(d\boldsymbol{y})\\ &=\int\int f(\boldsymbol{\Psi}(\boldsymbol{y}),\boldsymbol{y})\min[1,e^{H(\boldsymbol{y})-H(\boldsymbol{\Psi}(\boldsymbol{y}))}]\pi(d\boldsymbol{y}).\end{aligned}$$

The second part satisfies that

$$\begin{aligned}&\int\int f(\boldsymbol{x},\boldsymbol{x})\left(1-\min[1,e^{H(\boldsymbol{x})-H(\boldsymbol{\Psi}(\boldsymbol{x}))}]\right)\pi(d\boldsymbol{x})\\ &=\int\int f(\boldsymbol{y},\boldsymbol{y})\left(1-\min[1,e^{H(\boldsymbol{y})-H(\boldsymbol{\Psi}(\boldsymbol{y}))}]\right)\pi(d\boldsymbol{y}).\end{aligned}$$

By combining these two equations, we can see that the reversibility condition (18) holds. ■

In practice, one does not need to flip the momentum at the end of the second step, since it will be resampled in the first step in the next iteration. This was nevertheless required for showing the reversibility.

The following two examples illustrate the behaviour of Hamiltonian Monte Carlo.

Example 2 *In this example, we did some simulations for the simple 2D Gaussian distribution with covariance matrix* $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.98 \\ 0.98 & 1 \end{pmatrix}$. *We have used mass matrix* $\boldsymbol{M} = \boldsymbol{I}_2$, *and 25 leapfrog steps per iteration using stepsize* 0.25. *Figures 3 and 4 illustrate the behaviour of Hamiltonian Monte Carlo in this example, and compare it with random walk Metropolis.*

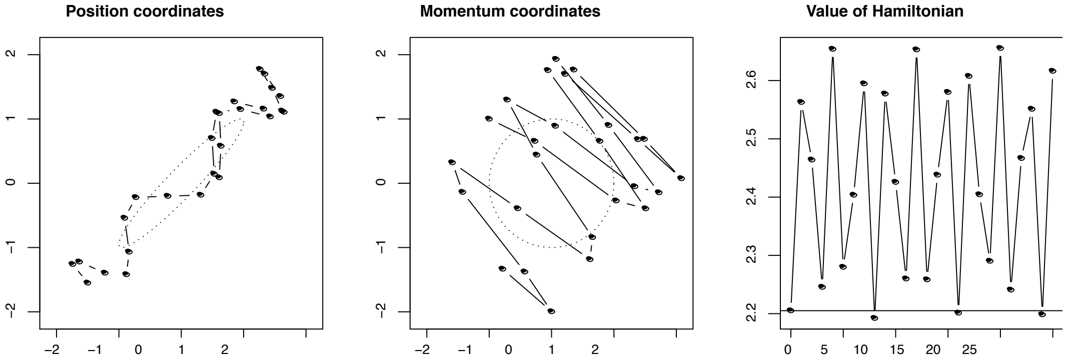


Figure 3: The position, momentum coordinates and the value of the Hamiltonian for 25 Leapfrog steps. As we can see, the Hamiltonian is kept approximately constant by the Leapfrog steps, hence the acceptance rate is close to 1.

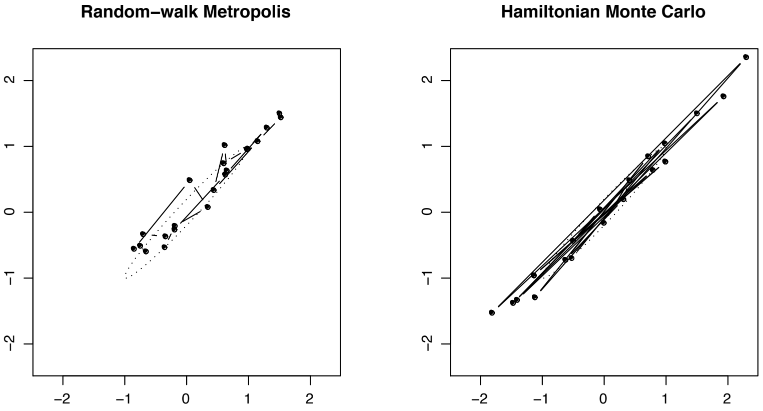


Figure 4: Twenty iterations of the random walk Metropolis method (20 updates per iteration) and the Hamiltonian Monte Carlo method (20 Leapfrog steps) for a highly correlated 2 dimensional Gaussian distribution. As we can see, Hamiltonian Monte Carlo is making much larger moves and mixes faster than random walk Metropolis.

Example 3 *In this example, we did some simulations for a 100 dimensional multivariate Gaussian distribution where the components are independent with standard deviations* $0.01, 0.02, \ldots, 1.00$. *The parameter* ϵ *was randomly chosen at each iteration uniformly from* $(0.0104, 0.0156)$, *and we selected* $L = 150$. *We compared HMC with the random walk Metropolis chain, counting 150 Metropolis updates per one iteration to be fair. The Figure 5 shows the last component (which had the largest standard deviation) for the two chains.*

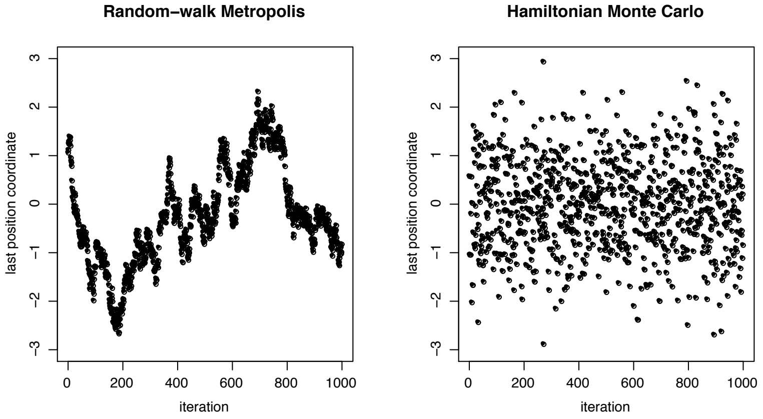


Figure 5: The last component in a 100 dimensional Gaussian distribution, random walk Metropolis versus Hamiltonian Monte Carlo. Hamiltonian Monte Carlo is mixing much faster.

5 Concentration of measure and high dimensional behaviour

As we have seen in the previous examples, HMC had performed much better than random walk Metropolis in high dimensions. To understand the reason for this, it is important to study the concentration of measure phenomenon that appears in high dimensions.

Suppose that $\boldsymbol{Z} \sim N(0, \boldsymbol{I}_d)$ is a d dimensional standard normal random vector. Then the Euclidean norm of $\boldsymbol{Z}$ satisfies that for every $t \geq 0$,

$$\boldsymbol{P}(|\|\boldsymbol{Z}\| - \sqrt{d}| \geq t) \leq C \exp\left(-\frac{t^2}{C}\right), \tag{19}$$

where C is an absolute constant independent of d. This means that with high probability, $\|\boldsymbol{Z}\| = \sqrt{d} + O(1)$, i.e. most of the probability is is concentrated in a thin layer around the sphere of radius $\sqrt{d}$. Note that here $H(\boldsymbol{z}) = \frac{\|\boldsymbol{z}\|^2}{2}$, and since $H(\boldsymbol{z})$ is preserved by the Hamiltonian, Hamiltonian dynamics moves around in circular arcs. See Figure 6 for an illustration of this.

In general, if the Hessian of the target potential satisfies that $\mu \boldsymbol{I}_d \preceq \nabla^2 U(\boldsymbol{x}) \preceq L\boldsymbol{I}_d$ for some $0 < \mu < L < \infty$ (strongly convex and smooth potential), and we let $H_{\min} := \inf_{\boldsymbol{z}} H(\boldsymbol{z})$, then it is possible to show that

$$\boldsymbol{P}\left(\left|\sqrt{H(\boldsymbol{z}) - H_{\min}} - \mathbb{E}\sqrt{H(\boldsymbol{z}) - H_{\min}}\right| \geq t\right) \leq C \exp\left(-\frac{t^2}{C}\right), \tag{20}$$

for some constant C that depends on μ and L but is independent of the dimension d. Hence the Hamiltonian is close to constant in the area of the space with high probability density, and therefore Hamiltonian dynamics can be very efficient in exploring this potentially complicated set automatically. In contrast to this, random walk Metropolis will need to take small moves of size $O(1)$ to keep acceptance rates reasonably high, and therefore it takes a long time to explore such high dimensional distributions.

References

[1] Duane, Simon and Kennedy, Anthony D and Pendleton, Brian J and Roweth, Duncan, Hybrid Monte Carlo, *Physics letters B*, vol. 195, no. 2, pp. 216-222, 1987.

[2] Neal, Radford M, MCMC using Hamiltonian dynamics, Handbook of Markov Chain Monte Carlo, vol. 2, no. 11, 2011

[3] Leimkuhler, Benedict and Reich, Sebastian, Simulating Hamiltonian dynamics, *Cambridge University Press*, vol. 14, 2004

[4] Betancourt, Michael, A conceptual introduction to Hamiltonian Monte Carlo, `https://arxiv.org/abs/1701.02434`

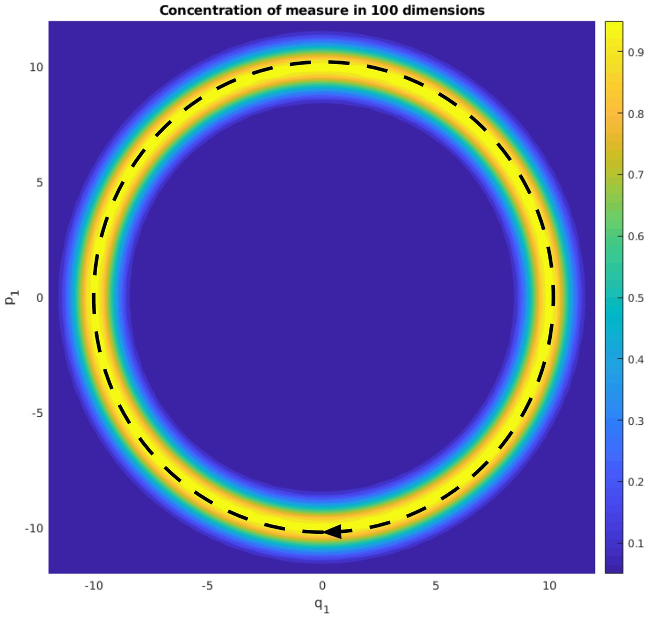


Figure 6: The density of the distribution π for a 100 dimensional standard Gaussian, when looking at the first components p_1, q_1 (other components have been integrated out). The dashed line shows a possible Hamiltonian path in this case.