Lecture 10:

Topics:

I) Hamiltonian Markov chain

II) Look ahead HMC.

Suppose d is some data vector drawn from some density $p(d|\theta)$ where $\theta = (\theta_1, \dots, \theta_n)$ is a parameter vector. Let $\pi(\theta)$ denote a prior density for θ.

Goal: Produce posterior samples

$$\theta^{(1)}, \dots, \theta^{(N)} \sim p(\theta|d) = \frac{p(d|\theta)p(\theta)}{p(d)}.$$

Problem: if the dimension n of $\theta \in \mathbb{R}^n$ is large then Metropolis-Hastings chains typically have low acceptance rates.

Hamiltonian Monte Carlo (HMC)

... an MCMC chain which attempts to solve the low acceptance rate problem when θ is high dimensional.

The main requirement for this method is the ability to compute

$$\nabla_\theta \log p(\theta|d) \quad \& \quad p(\theta|d)$$

quickly (where $p(\theta|d)$ only needs to be computed up to a unknown normalizing factor).

To define the HMC sample, start by defining a momentum vector

$$p := (p_1, \dots, p_n) \in \mathbb{R}^n$$

with corresponding mass

$$m := (m_1, \dots, m_n) \in \mathbb{R}^n \quad \text{s.t.} \quad m_i > 0.$$

Now define the Hamiltonian as follows

$$\begin{aligned} H(\theta, p) &:= -\log p(\theta|d) + \sum_{i=1}^{n} \frac{|p_i|^2}{2m_i} \\ &= -\log p(\theta|d) - \log \pi(p) + c \\ &= -\log \left[p(\theta|d)\pi(p) \right] + c \end{aligned}$$

where $\pi(p)$ is the density of $p \sim N(0, \text{diag}(m))$.

You can think of $-\log \pi(p)$ as the kinetic energy of the system (θ, p) and $-\log p(\theta|d)$ as potential energy.

If we are thinking about (θ, p) as describing a physical system at time $t=0$, then we can evolve the system, producing (θ^t, p^t) for $t \in [0, \infty)$, with Hamiltonian dynamics:

$$(H1) \quad \frac{d\theta^t}{dt} = \nabla_p H(\theta^t, p^t)$$

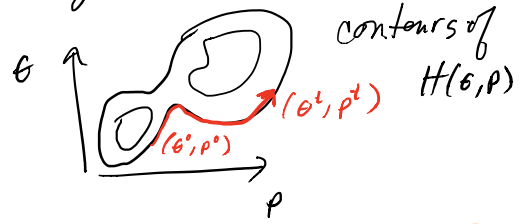
$$(H2) \quad \frac{dp^t}{dt} = -\nabla_\theta H(\theta^t, p^t)$$

Under this evolution $H(\theta^t, p^t)$ is preserved

$H(\theta^t, p^t) \equiv H(\theta^0, p^0) \quad \forall t \in [0, \infty)$ since

$$\begin{aligned} \frac{d}{dt} H(\theta^t, p^t) &= \nabla_\theta H(\theta^t, p^t) \cdot \frac{d\theta^t}{dt} + \nabla_p H(\theta^t, p^t) \frac{dp^t}{dt} \\ &= 0. \end{aligned}$$

$\therefore$ (θ^t, ρ^t) moves along a contour of the log likelihood given by $H(\theta,\rho)+c$



One can also show that (H1) & (H2) preserve volumn in "phase space $(\theta,\rho)\in\mathbb{R}^n\times\mathbb{R}^n$" meaning that for any set $\Omega\subset\mathbb{R}^n\times\mathbb{R}^n$ one has

$$\text{vol}(\Omega) = \text{vol}\left(\mathcal{H}^t(\Omega)\right)$$

where

$$\mathcal{H}^t(\Omega) := \left\{ \mathcal{H}^t(\theta,\rho)\in\mathbb{R}^n\times\mathbb{R}^n \;\middle|\; (\theta,\rho)\in\Omega \right\}$$

$\mathcal{H}^t(\theta,\rho) = (\theta^t,\rho^t)$ evolved by (H1)(H2) with initial condition $(\theta^0,\rho^0) := (\theta,\rho)$.

To see this simply note that $\mathcal{H}^t(\theta,\rho)$ is a div free evolution.

i.e. that

$$\text{div}\left((\nabla_\rho H, -\nabla_\theta H)^T \right) = \sum_{i=1}^n \frac{\partial^2}{\partial\theta_i\partial\rho_i} H(\theta,\rho) - \sum_{i=1}^n \frac{\partial^2}{\partial\rho_i\partial\theta_i} H(\theta,\rho)$$

$$= 0$$

Why is this useful?

Let $X \equiv (\theta^0,\rho^0)$ be a random vector with some density $\rho_X(x)$.

Transform $X \longmapsto Y$ as follows

$$Y = \mathcal{H}^t(X) = (\theta^t,\rho^t) \text{ under (H1)(H2)}.$$

the change of variables formula says

$$\rho_X(x)\,\partial x = \rho_Y(y)\,\partial y$$

$\therefore$

$$\rho_X(x) = \rho_Y(y)\left|\frac{\partial y}{\partial x}\right|$$

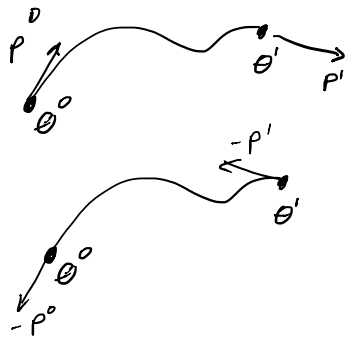
$\leftarrow$ jacobian

$$= \rho_Y(\mathcal{H}^t(x))\left|\frac{\partial \mathcal{H}^t(x)}{\partial x}\right|$$

$= 1$ by vol preserving property

i.e. $\rho_X(\theta^0,\rho^0) = \rho_Y(\theta^t,\rho^t)$

Finally we note that (H1) & (H2) is reversable in that flipping the sign on momentum ρ runs the chain in reverse.



Let $s \equiv (\theta, p)$ denote general state vectors. Let L be the operator on state vectors defined by evolving (H1) & (H2) to some fixed time t.

$$\therefore \quad Ls \equiv L(\theta,p)^T := \mathcal{H}^t(\theta,p)$$

Also let F denote the "momentum sign flip operator" so that

$$Fs \equiv F(\theta,p)^T := (\theta, -p).$$

Note that our previous results on (H1) & (H2) imply

$$L^{-1}s = FLFs \quad \text{(Reversability)}$$

$$\left|\frac{\partial Ls}{\partial s}\right| = 1 \quad \text{(vol preserving)}$$

$$H(Ls) = H(s) \quad \text{(conservation of energy)}$$

we also have

$$F^{-1}s = Fs$$

$$\left|\frac{\partial Fs}{\partial s}\right| = 1 \quad \text{(vol preserving)}$$

$$H(Fs) = H(s).$$

Finally define the "momentum Randomizing operator" R_β (parameter $\beta \in [0,1]$) as follows

$$R_\beta s \equiv R_\beta(\theta,p)$$
$$:= (\theta, p\sqrt{1-\beta} + \sqrt{\beta}\, p^*)$$
$$p^* \sim N(0, \mathrm{diag}(m))$$

Now we can construct a markov chain $s^{(1)}, s^{(2)}, \ldots$ for state vectors

$$s^{(i)} \equiv (\theta^{(i)}, p^{(i)})$$

as follows.

Hamiltonian Markov Chain

for $i = 1, 2, \ldots$

(this is basically a proposal.)

$$\alpha := \min\left(\frac{\exp(-H(FLs^{(i)}))}{\exp(-H(s^{(i)}))}, 1\right)$$

$$\tilde{s} := \begin{cases} FLs^{(i)} & \text{if } u \le \alpha \\ s^{(i)} & \text{if } u > \alpha \end{cases}$$

$$s^{(i+1)} := R_\beta F \tilde{s}$$

(then an additional random walk step)

end.

The key for showing this works (i.e. has invariant density $e^{-H(s)} = e^{-H(\theta,p)}$) is the Reversability of L & the vol preserving of L & F.

In this form it is hard to see what is going on. We can simplify the above algorithm when $\beta = 1$. Notice the momentum components of $s^{(i)}$ are eventually discarded so we can write the algorithm in terms of $\theta^{(i)}$ alone.

Hamiltonian Monte Carlo $\beta = 1$

for $i = 1, 2, \ldots$

Simulate $p^* \sim N(0, \text{diag}(m))$

$\tilde{s} := (\theta^{(i)}, p^*)$

$\alpha := \min\left(\frac{\exp(-H(L\tilde{s}))}{\exp(-H(\tilde{s}))}, 1\right)$

$\theta^{(i+1)} := \begin{cases} (L\tilde{s})_\theta & \text{if } u \le \alpha; \\ \tilde{s}_\theta & \text{if } u > \alpha. \end{cases}$

end

where s_θ extracts the parameter coordinates from s.

This is basically RWF $\tilde{s}$ from the previous step

Using the exact Hamiltonian evolution operator L, since $H(L\tilde{s}) = H(\tilde{s})$ α is always 1 so this mcmc always accepts the proposal.

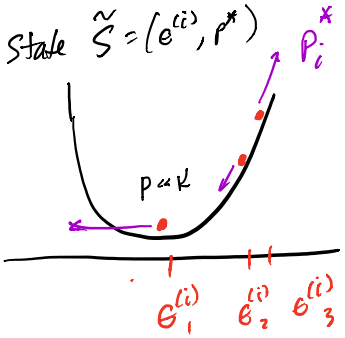
To get an idea of what is going on expand out the Hamiltonian

$$H(\tilde{s}) = H(\theta^{(i)}, p^*)$$

$$= -\log p(\theta^{(i)}|d) + \sum_{i=1}^{n} \frac{|p_i^*|^2}{2m_i}$$

Potential energy (T) — Kinetic energy (K)

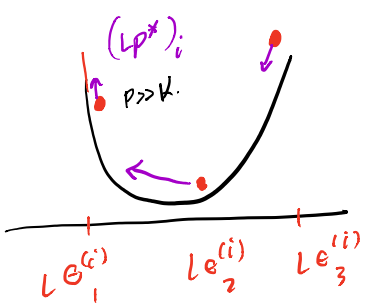
This will be a $\frac{\chi^2_n}{2}$ which is how $-\log p(\theta^{(i)}|d)$ should approximately vary when $\theta^{(i)} \sim p(\theta|d)$



The Hamiltonian is evolved with L which where Kinetic & potential energy are exchanged.

$$H(L\tilde{s}) = -\log p(L\theta^{(i)}|d) + \sum_{i=1}^{n} \frac{|Lp_i^*|^2}{2m_i}$$

$$L\tilde{s} = (L\theta^{(i)}, Lp^*)$$



This changes the Potential energy $-\log p(\theta|d)$ by an amount expected from samples $\theta \sim p(\theta|d)$

Showing detailed balance

Consider a Markov chain $X^{(1)}, X^{(2)}, \dots$

Let T be a random operator s.t. the $i+1^{st}$ step can be written

$$X^{(i+1)} = TX^{(i)}$$

↖ the random seeds is independently drawn for each i.

The notion of detailed balance for transition operator T & a measure π is simply that when $X \sim \pi$ then $(X, TX) \overset{\mathcal{D}}{=} (TX, X)$.

This clearly implys $X \overset{\mathcal{D}}{=} TX$ (when $X \sim \pi$) so detailed balance implies π is an invariant measure for the chain.

Notice that if $T = T_N \cdots T_1$, i.e. there are N intermediate steps in each transition $X^{(i)} \to X^{(i+1)}$:

$$X^{(i)} \to T_1 X^{(i)} \to T_2 T_1 X^{(i)} \to \cdots \to TX^{(i)} = X^{(i+1)}$$

Then to prove π is an invariant measure it is sufficient to show each T_i satisfies detailed balance w.r.t π (follows since each application of T_i is invariant to π).

For the HMC each transition $s^{(i)} \to s^{(i+1)}$ is composed of two steps

$$s^{(i)} \to \tilde{s} \to s^{(i+1)}$$

write this abstractly as

$$s \to Ts \to R_{\beta F} Ts$$

The goal of this section is to illustrate part of the proof that T_1 satisfies detailed balance for the measure $e^{-H(s)}$ and show where the vol preserving nature of F & L and reversibility $L^{-1} = FLF$ are used.

Let $s \sim e^{-H(s)}$ & consider the two joint R.V.s

$$(s, Ts) \quad \& \quad (Ts, s)$$

we wish to show they have the same distribution where

$$Ts := \begin{cases} FLs & \text{with prob } \alpha(FLs \mid s) \\ s & \text{with prob } 1 - \alpha(FLs \mid s) \end{cases}$$

$$\alpha(x \mid y) := \min\left(\frac{e^{-H(x)}}{e^{-H(y)}}, 1\right)$$

Think of proving detail balance as a likelihood inference problem

Given two states X, Y in phase space

we ask the question: is it equally likely that

$$s = X \text{ or } s = Y$$
$$Ts = Y \qquad Ts = X$$

assuming $s \sim e^{-H(s)}$.

To satisfy detailed balance you need any likelihood discrepancy btwn

$$e^{-H(X)} \quad \& \quad e^{-H(Y)}$$

need to be canceled out by the likelihoods of the transitions

$$T(X) = Y \quad \& \quad T(Y) = X$$

Let P be the prob measure for (X,Y) under the assumption $(X,Y) := (S, TS)$ when $S \sim e^{-H(s)}$. We wish to show

$$P(X \in A, Y \in B) = P(Y \in A, X \in B).$$

we will only do part of the proof, assuming $A \cap B = \emptyset$. This implies TS must have resulted in $TS = FLS$

$\therefore P(X \in A, Y \in B)$ ← if $A \cap B = \emptyset$ then $T = FL$ & $Y = FLX$

$= P(X \in A, FLX \in B, T = FL)$

$= P(X \in A \cap (FL)^{-1}B, T = FL)$

since $(FL)^{-1} = L^{-1}F = FLFF = L$ by reversibility !!!

$= P(X \in A \cap FLB, T = FL)$

$= \int_{A \cap FLB} \alpha(FLx|x)\, e^{-H(x)}\, dx$ ← $= P(T = FL \mid X = x)$

$= \int_{A \cap FLB} \min\left(e^{-H(FLx)}, e^{-H(x)}\right) dx$

change variables $y = FLx$, $dy = dx$ ∴ by vol preserving of FL !!!

$= \int_{FLA \cap B} \alpha(FLy|y)\, e^{-H(y)}\, dy$

$= P(X \in B, Y \in A)$

$= P(Y \in A, X \in B)$ ∎

Leap frog discretizing L

Recall that L is defined to be the operator on state vectors s which evolves (H1) & (H2) to some fixed time t.

The key properties of Reversibility & vol preserving establish that the HMC algorithm has invariant density given by $e^{-H(s)}$.

It turns out that there exists a discrete approximation to L, denoted by $L^{\varepsilon,m}$, that still satisfies Reversibility & vol preservation, but for which $H(s)$ is not preserved (so $H(s) \neq H(L^{\varepsilon m} s)$) as is the case for L.

$L^{\varepsilon m}$ is called leapfrog, or Störmer-Verlet, integration and is defined as follows

function $L^{\varepsilon m}(s)$

  set $s^0 := s$ and $(\theta^0, p^0) = s^0$.

  set $\Lambda = \mathrm{diag}(m_1, \dots, m_n)$

  for $k = 0, 1, \dots, M$

    $\theta^{k+1} := \theta^k + \varepsilon \Lambda^{-1}\left[p^k - \frac{\varepsilon}{2} \nabla_\theta H(\theta^k, p^k)\right]$

    $p^{k+1} := p^k - \frac{\varepsilon}{2}\left[\nabla_\theta H(\theta^k, p^k) + \nabla_\theta H(\theta^{k+1}, p^k)\right]$

    end

  return (θ^{M+1}, p^{M+1})

end

Note: the updates for θ^{k+1}, p^{k+1} are usually written as

$$\tilde{p} := p_k - \frac{\varepsilon}{2}\nabla_\theta H(\theta^k, p^k)$$

$$\theta^{k+1} := \theta^k + \varepsilon \nabla_p H(\theta^k, \tilde{p})$$

$$p^{k+1} := \tilde{p} - \frac{\varepsilon}{2}\nabla_\theta H(\theta^{k+1}, \tilde{p})$$

which more closely resemble discrete Hamiltonian updates (but are exactly equivalent to the algorithm given).

Now the Discrete HMC simply replaces L with $L^{\varepsilon,M}$.

The main difference using $L^{\varepsilon,M}$ in place of L is that acceptance rate α will no longer be exactly 1 since the Hamiltonian is only approximately preserved.

Indeed, ε & M play the roll of tuning parameters where larger ε and/or larger M will result in lower acceptance rates ... but give larger moves when accepted.

One of the main drawbacks with HMC is that it can be difficult to tune.

Tuning Parameters

* $\beta > 0$ can be useful (see Culpepper 2011)

* The typical suggestion is to set $\mathrm{diag}(M)$ [or a general Λ where $p \sim N(0,\Lambda)$] to the inverse of the posterior cov matrix of θ.

* Then ε & M are chosen so that updating with $L^{\varepsilon,M}$ gives good acceptance.

CMB application

Taylor, Ashdown & Hobson used HMC to sample from $\underbrace{P(T, C_\ell^{TT} | d)}_{\equiv P(\theta|d)}$ using the fact that

$$-\nabla_\theta \log P(\theta|d) \equiv \begin{pmatrix} -\nabla_{T(\hat{n})} \log P(T, C_\ell^{TT}|d) \\ -\nabla_{C_\ell^{TT}} \log P(T, C_\ell^{TT}|d) \end{pmatrix}$$

where

$$-\nabla_{T(\hat{n})} \log P(T, C_\ell^{TT}|d) = -(\Sigma + N)^{-1}(d - \vec{T}) + \Sigma^{-1}\vec{T}$$

$$-\nabla_{C_\ell^{TT}} \log P(T, C_\ell^{TT}|d) = \left(\frac{2\ell+1}{2}\right)\frac{1}{C_\ell^{TT}}\left(1 - \frac{\sigma_\ell^T}{C_\ell^{TT}}\right)$$

when $d_{\ell m} = T_{\ell m} + n_{\ell m}$, one uses a uniform prior on C_ℓ^{TT} &

$$\sigma_\ell^T := \frac{1}{2\ell+1} \sum_{m=-\ell}^{\ell} |T_{\ell m}|^2.$$

Look Ahead Hamiltonian Monte Carlo (LAHMC)

The detailed balance property of HMC suggest that even though one gets high acceptance rates there is still random walk behavior in HMC leading to slow mixing.

Sohl-Dickstein et. al. developed a similar chain to HMC with the same stationary distribution but which doesn't satisfy detailed balance & can have a faster mixing rate.

LAHMC Produces $s^{(1)}, s^{(2)}, \ldots$ in phase space

Fix $K \in \{1, 2, \ldots\}$, $0 \le \beta \le 1$ and let L denote the leapfrog operator $L^{\varepsilon, m}$ where $L^j := \underbrace{L L \cdots L}_{j \text{ times}}$.

Define

$$\alpha_1(s) := \min\left[\frac{e^{-H(Fs)}}{e^{-H(s)}}, 1 \right]$$

$$\alpha_k(s) := \min\left[\frac{e^{-H(FL^k s)}}{e^{-H(s)}} \left(1 - \sum_{i=1}^{k-1} \alpha_i(FL^k s)\right), 1 - \sum_{i=1}^{k-1} \alpha_i(s) \right]$$

for $k = 2, \ldots, K$ and

$$\alpha_{K+1}(s) := 1 - \sum_{k=1}^{K} \alpha_k(s).$$

for $i = 1, 2, \ldots$

$$\tilde{s} := \begin{cases} FL^1 s^{(i)} & \text{with prob } \alpha_1(s^{(i)}) \\ FL^2 s^{(i)} & \text{with prob } \alpha_2(s^{(i)}) \\ \vdots & \\ FL^K s^{(i)} & \text{with prob } \alpha_K(s^{(i)}) \\ s^{(i)} & \text{with prob } \alpha_{K+1}(s^{(i)}) \end{cases}$$

$$s^{(i+1)} = R_\beta F \tilde{s}$$

end

To see why LAHMC works lets again take a look at the transition

$$s^{(i)} \xrightarrow{T} \tilde{s}$$

and show $e^{-H(s)}$ is an invariant measure.

So suppose $s \sim e^{-H(s)}$

$$\therefore \quad Ts = \begin{cases} FL^1 s & \text{w/prob } \alpha_1(s) \\ \vdots & \vdots \\ FL^K s & \text{w/prob } \alpha_K(s) \\ s & \text{w/prob } \alpha_{K+1}(s) \end{cases}$$

$$\therefore P(T_S \in A) = \int \underbrace{P(T_S \in A \mid s)}\, e^{-H(s)} ds$$

$$= \sum_{k=1}^{K} P(FL^k S \in A \mid s)\, \alpha_k(s) + P(FS \in A \mid s)\, \alpha_{K+1}(s)$$

$$= \sum_{k=1}^{K} \int_{(FL^k)^{-1}A} \alpha_k(s) e^{-H(s)} ds + \int_A \alpha_{K+1}(s) e^{-H(s)} ds$$

$$= \sum_{k=1}^{K} \int_A \alpha_k(FL^k S) e^{-H(FL^k S)} ds + \int_A \alpha_{K+1}(s) e^{-H(s)} ds$$

Since $(FL^k)^{-1} = FL^k$ is vol preserving

Now

$$\alpha_k(FL^k S) e^{-H(FL^k S)} \qquad = S$$

$$= \min\left[e^{-H(s)} \left(1 - \sum_{i=1}^{k-1} \alpha_i(\overbrace{FL^k FL^k}^{} s)\right), \right.$$
$$\left. e^{-H(FL^k S)} \left(1 - \sum_{i=1}^{k-1} \alpha_i(FL^k S)\right)\right]$$

$$= \alpha_k(s)\, e^{-H(s)}$$

and

$$\therefore P(T_S \in A) = \sum_{k=1}^{K} \int_A \alpha_k(s) e^{-H(s)} ds + \int_A \alpha_{K+1}(s) e^{-H(s)} ds$$

$$= \int_A \underbrace{(\alpha_1(s) + \cdots + \alpha_{K+1}(s))}_{=1} e^{-H(s)} ds$$

$\therefore T_S \sim e^{-H(s)}$ as was to be shown.