MAST90125: Bayesian Statistical learning

Lecture 19: Hamiltonian Monte Carlo

Feng Liu and Guoqi Qian

THE UNIVERSITY OF MELBOURNE

What do we know about MCMC so far

- We introduced MCMC methods (Metropolis-Hastings and Gibbs). Remember these methods were used to make draws from the posterior distribution, $p(\boldsymbol{\theta}|\mathbf{y})$ when we cannot determine $p(\boldsymbol{\theta}|\mathbf{y})$ analytically.
- In the process, we noted
  - that MCMC methods produce dependent samples, which reduces the effective sample size.
  - that MCMC methods can take a long time to converge to the posterior distribution.

Hamiltonian Monte Carlo

- For the remainder of this lecture, we will discuss Hamiltonian (or hybrid) Monte Carlo.
  - This is a technique designed to reduce correlation between successive iterations. Consequently, HMC should move more rapidly towards the target distribution.
  - In fact, we have already used Hamiltonian Monte Carlo. The software `Stan` uses Hamiltonian Monte Carlo to fit models in a Bayesian framework. We have already used `Stan` in Lecture 18 Rscript. Despite this we will develop an R program for HMC.

Theory behind Hamiltonian Monte Carlo

- Concept of conservation of energy of a particle system:

$$H(t) = U(\mathbf{q}(t)) + K(\mathbf{p}(t)),$$

where $H(t)$ is the Hamiltonian, $U(\mathbf{q}(t))$ is the potential energy and $K(\mathbf{p}(t))$ is the kinetic energy at time t, position $\mathbf{q}(t)$ and momentum $\mathbf{p}(t)$.

- As energy is conserved, we know that $dH(t)/dt = 0$, which implies that

$$0 = \frac{dH(t)}{dt} = \frac{\partial H}{\partial \mathbf{q}'} \frac{d\mathbf{q}(t)}{dt} + \frac{\partial H}{\partial \mathbf{p}'} \frac{d\mathbf{p}(t)}{dt}$$

which has solutions

$$\frac{d\mathbf{q}(t)}{dt} = +\frac{\partial H}{\partial \mathbf{p}}, \quad \frac{d\mathbf{p}(t)}{dt} = -\frac{\partial H}{\partial \mathbf{q}}.$$

Theory behind Hamiltonian Monte Carlo

- The negative log-posterior $-\log(p(\boldsymbol{\theta}|\mathbf{y}))$ is deemed a potential energy evaluated at random position $\boldsymbol{\theta}$. But then, what is the momentum?
- For the momentum, create an auxiliary variable ϕ drawn from distribution $p(\phi|\boldsymbol{\theta})$. Then the negative log density, $-\log(p(\phi|\boldsymbol{\theta}))$ is the kinetic energy, so the Hamiltonian becomes,

$$H(t) = -\log(p(\boldsymbol{\theta}^{(t)}|\mathbf{y})) - \log(p(\phi^{(t)}|\boldsymbol{\theta}^{(t)})).$$

Theory behind Hamiltonian Monte Carlo

- Then the question becomes, what distribution should we use for ϕ? While the choice is flexible, the formula of kinetic energy will be indicative,

$$\frac{m\mathbf{v}'\mathbf{v}}{2} = \frac{m\mathbf{v}'m\mathbf{v}}{2m} = \mathbf{p}'\frac{1}{2m}\mathbf{p}, \quad \text{as the momentum } \mathbf{p} = m\mathbf{v}.$$

If we let our defined kinetic energy equal $\mathbf{p}'\frac{1}{2m}\mathbf{p}$, then we can find that

$$-\log(p(\phi^{(t)}|\theta^{(t)})) = \mathbf{p}'\frac{1}{2m}\mathbf{p}$$

Theory behind Hamiltonian Monte Carlo

- Then the question becomes, what distribution should we use for ϕ? While the choice is flexible, the formula of kinetic energy will be indicative,

$$\frac{m\mathbf{v}'\mathbf{v}}{2} = \frac{m\mathbf{v}'m\mathbf{v}}{2m} = \mathbf{p}'\frac{1}{2m}\mathbf{p}, \quad \text{as the momentum } \mathbf{p} = m\mathbf{v}.$$

  If we let our defined kinetic energy equal $\mathbf{p}'\frac{1}{2m}\mathbf{p}$, then we can find that

$$-\log(p(\phi^{(t)}|\boldsymbol{\theta}^{(t)})) = \mathbf{p}'\frac{1}{2m}\mathbf{p} \rightarrow p(\phi^{(t)}|\boldsymbol{\theta}^{(t)}) = e^{-(\mathbf{p}-\mathbf{0})'(2m)^{-1}(\mathbf{p}-\mathbf{0})},$$

  strongly suggesting choose $p(\phi) = \mathcal{N}(\mathbf{0}, \mathbf{M})$, where $\mathbf{M}$ is the 'mass' matrix.
- Note that generating $\boldsymbol{\theta}^{(t)}$'s from the posterior $p(\boldsymbol{\theta}|\mathbf{y})$ now becomes generating $(\boldsymbol{\theta}^{(t)}, \phi^{(t)})$'s to stabilize $H(t)$. This can be achieved by a Monte Carlo method.

Implementing Hamiltonian Monte Carlo

- Having determined the 'potential' and 'kinetic' energies, we need the derivatives to implement the Monte Carlo method. These are

$$\frac{\partial H(t)}{\partial \phi} = \frac{\partial\{-\log(p(\boldsymbol{\theta}^{(t)}|\mathbf{y})) - \log(p(\phi^{(t)}|\boldsymbol{\theta}^{(t)}))\}}{\partial \phi} = -\frac{\partial \log(p(\phi^{(t)}|\boldsymbol{\theta}^{(t)}))}{\partial \phi}$$
$$\frac{\partial H(t)}{\partial \theta} = \frac{\partial\{-\log(p(\boldsymbol{\theta}^{(t)}|\mathbf{y})) - \log(p(\phi^{(t)}|\boldsymbol{\theta}^{(t)}))\}}{\partial \boldsymbol{\theta}} = -\frac{\partial \log(p(\boldsymbol{\theta}^{(t)}|\mathbf{y}))}{\partial \boldsymbol{\theta}} - \frac{\partial \log(p(\phi^{(t)}|\boldsymbol{\theta}^{(t)}))}{\partial \boldsymbol{\theta}}.$$

- However if we use the standard assumption that $p(\phi) = \mathcal{N}(\mathbf{0}, \mathbf{M}) = (2\pi)^{-k/2} \det(\mathbf{M})^{-1/2} e^{-\phi' \mathbf{M}^{-1} \phi/2}$, the term $\frac{\partial \log(p(\phi^{(t)}|\boldsymbol{\theta}^{(t)}))}{\partial \boldsymbol{\theta}}$ disappears and $\log(p(\phi))$ becomes,

$$-0.5k \log(2\pi) - 0.5 \log(\det(\mathbf{M})) - 0.5\phi' \mathbf{M}^{-1} \phi$$

Implementing Hamiltonian Monte Carlo

- Having made these decisions, the derivatives of interest are
  - $\frac{\partial H}{\partial \phi} = \mathbf{M}^{-1}\phi$
  - $\frac{\partial H}{\partial \boldsymbol{\theta}} = -\frac{d \log(p(\boldsymbol{\theta}|\mathbf{y}))}{d\boldsymbol{\theta}} = -\frac{d\{\log(p(\boldsymbol{\theta}, \mathbf{y})) - \log(p(\mathbf{y}))\}}{d\boldsymbol{\theta}} = -\frac{d \log(p(\boldsymbol{\theta}, \mathbf{y}))}{d\boldsymbol{\theta}}$
- Now the question is how to generate ϕ, $\boldsymbol{\theta}$ that satisfy the Hamiltonian equations. Since we are working with chains, we will have already drawn $\boldsymbol{\theta}^{(t)}, \phi^{(t)}$. So just draw $\boldsymbol{\theta}^{(t+\epsilon)}, \phi^{(t+\epsilon)}$ such that,

$$\frac{d\boldsymbol{\theta}}{dt} = \frac{\boldsymbol{\theta}^{(t+\epsilon)} - \boldsymbol{\theta}^{(t)}}{\epsilon} = \frac{\partial H}{\partial \phi} = \mathbf{M}^{-1}\phi$$

$$\frac{d\phi}{dt} = \frac{\phi^{(t+\epsilon)} - \phi^{(t)}}{\epsilon} = -\frac{\partial H}{\partial \boldsymbol{\theta}} = \frac{d \log(p(\boldsymbol{\theta}, y))}{d\boldsymbol{\theta}}$$

Steps of Hamiltonian Monte Carlo

- The following 'leapfrog' algorithm is for updating $\boldsymbol{\theta}$ (and $\boldsymbol{\phi}$).
  - Assume we are in state $t-1$. In conjunction to $\boldsymbol{\theta}^{(t-1)}$, sample $\boldsymbol{\phi}^{(t-1)}$ from $p(\boldsymbol{\phi})$.
  - For $i = 1, \ldots, L$
    - Set $\boldsymbol{\phi}^{(t-1+(i-1/2)\epsilon)} = \boldsymbol{\phi}^{(t-1+(i-1)\epsilon)} + \frac{\epsilon}{2}\frac{d\log(p(\boldsymbol{\theta},\mathbf{y}))}{d\boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t-1+(i-1)\epsilon)}}$
    - Set $\boldsymbol{\theta}^{(t-1+i\epsilon)} = \boldsymbol{\theta}^{(t-1+(i-1)\epsilon)} + \epsilon\mathbf{M}^{-1}\boldsymbol{\phi}^{(t-1+(i-1/2)\epsilon)}$
    - Set $\boldsymbol{\phi}^{(t-1+i\epsilon)} = \boldsymbol{\phi}^{(t-1+(i-1/2)\epsilon)} + \dfrac{\epsilon}{2}\dfrac{d\log(p(\boldsymbol{\theta},\mathbf{y}))}{d\boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t-1+i\epsilon)}}$
  - Label $\boldsymbol{\phi}^{(t)} = \boldsymbol{\phi}^{(t-1+L\epsilon)}, \boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1+L\epsilon)}$ and calculate

$$r = \frac{p(\boldsymbol{\theta}^{(t)}|\mathbf{y})p(\boldsymbol{\phi}^{(t)})}{p(\boldsymbol{\theta}^{(t-1)}|\mathbf{y})p(\boldsymbol{\phi}^{(t-1)})} = \frac{\frac{p(\boldsymbol{\theta}^{(t)},\mathbf{y})}{p(\mathbf{y})}p(\boldsymbol{\phi}^{(t)})}{\frac{p(\boldsymbol{\theta}^{(t-1)},\mathbf{y})}{p(\mathbf{y})}p(\boldsymbol{\phi}^{(t-1)})} = \frac{p(\boldsymbol{\theta}^{(t)},\mathbf{y})p(\boldsymbol{\phi}^{(t)})}{p(\boldsymbol{\theta}^{(t-1)},\mathbf{y})p(\boldsymbol{\phi}^{(t-1)})}$$

  - Set $\boldsymbol{\theta}^{(t)} = \begin{cases} \boldsymbol{\theta}^{(t)} & \text{with probability } \min(r,1) \\ \boldsymbol{\theta}^{(t-1)} & \text{otherwise} \end{cases}$

Comments on Hamiltonian Monte Carlo algorithm

- By splitting the updating of ϕ into half-steps, we ensure the symmetry of the algorithm. To undo the leapfrog steps, just replace ϕ with $-\phi$ as shown below.
- For $i = L, \ldots, 1$
  - Set $(-\phi)^{(t-1+(i-1/2)\epsilon)} = (-\phi)^{(t-1+i\epsilon)} + \frac{\epsilon}{2} \frac{d \log(p(\boldsymbol{\theta}, y))}{d\boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t-1+i\epsilon)}}$
  - Set $\boldsymbol{\theta}^{(t-1+(i-1)\epsilon)} = \boldsymbol{\theta}^{(t-1+i\epsilon)} + \epsilon \mathbf{M}^{-1}(-\phi)^{(t-1+(i-1/2)\epsilon)}$
  - Set $(-\phi)^{(t-1+(i-1)\epsilon)} = (-\phi)^{(t-1+(i-1/2)\epsilon)} + \frac{\epsilon}{2} \frac{d \log(p(\boldsymbol{\theta}, y))}{d\boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t-1+(i-1)\epsilon)}}$

Comments on Hamiltonian Monte Carlo algorithm

- This means the proposed conditional distributions are $J(\boldsymbol{\theta}^{(t)}, \phi^{(t)} | \boldsymbol{\theta}^{(t-1)}, \phi^{(t-1)}) = p(\phi^{(t-1)})$ and $J(\boldsymbol{\theta}^{(t-1)}, \phi^{(t-1)} | \boldsymbol{\theta}^{(t)}, \phi^{(t)}) = p(-\phi^{(t-1)})$. Moreover as $\phi \sim \mathcal{N}(\mathbf{0}, \mathbf{M})$, we know

$$p(\phi) = \frac{e^{-\phi' \mathbf{M}^{-1} \phi/2}}{(2\pi)^{k/2} \det(\mathbf{M})^{1/2}} = \frac{e^{-(-\phi)' \mathbf{M}^{-1} (-\phi)/2}}{(2\pi)^{k/2} \det(\mathbf{M})^{1/2}} = p(-\phi).$$

  Hence Hamiltonian Monte Carlo is a special case of a Metropolis-hasting algorithm (with a symmetry conditional distribution).

- Typically ϵ, L are chosen such that $\epsilon \times L = 1$ and L is an integer.

- According to theory, the optimal acceptance rate of a HMC algorithm should be ≈ 65 %, compared to ≈ 23 % for a Metropolis algorithm in a multi-dimensional problem.

Comments on Hamiltonian Monte Carlo algorithm

- **Very important**: We are not interested in updating ϕ, only $\boldsymbol{\theta}$. Hence at each state t, before starting the leapfrog steps, we sample $\phi^{(t)}$ from the prior, and not conditional on $\phi^{(t-1)}$.

- This means that while we assume the Hamiltonian is constant in sub-states $t-1+i\epsilon$; $1,\ldots,L$, we allow the Hamiltonian to change between states $t-2$, $t-1$, $t,\ldots$. If we did not allow the Hamiltonian to move between states, we would implicitly enforce bounds on $\log(p(\boldsymbol{\theta}, y))$ that would prevent full exploration of the posterior density.

Example of the Hamiltonian Monte Carlo algorithm

- To demonstrate HMC, we will look at the logistic regression example. As a reminder, this was

$$\Pr(y_i|p_i) = \mathrm{Bin}(n_i, p_i) \quad \log(p_i/(1-p_i)) = \mathbf{x}_i'\beta \quad p(\beta) \propto 1.$$

- As this is an example of a generalised linear model, the likelihood (and joint distribution, since $p(\beta) \propto 1$) we will work with is,

$$\Pr(\mathbf{y}|\beta) = \prod_{i=1}^{N} \binom{n_i}{y_i} e^{(\mathbf{x}_i'\beta)y_i}(1 + e^{(\mathbf{x}_i'\beta)})^{-n_i}$$

Example of the Hamiltonian Monte Carlo algorithm

- In order to implement Hamiltonian Monte Carlo, we need the derivative of the log joint distribution. The steps required to find this are,

$$
\begin{aligned}
\log(p(\boldsymbol{\beta}, \mathbf{y})) &= \log(\Pr(y|\boldsymbol{\beta})) + \log(p(\boldsymbol{\beta})) = \log(\Pr(y|\boldsymbol{\beta})) \quad \text{as } p(\boldsymbol{\beta}) \propto 1 \\
&= \sum_{i=1}^{N} \log\left(\binom{n_i}{y_i}\right) + \sum_{i=1}^{N} y_i(\mathbf{x}_i'\boldsymbol{\beta}) - \sum_{i=1}^{N} n_i \log(1 + e^{(\mathbf{x}_i'\boldsymbol{\beta})}) \\
\frac{d\log(p(\boldsymbol{\beta}, \mathbf{y}))}{d\boldsymbol{\beta}_j} &= \sum_{i=1}^{N} y_i \mathbf{x}_{ij} - \sum_{i=1}^{N} n_i \frac{\mathbf{x}_{ij} e^{(\mathbf{x}_i'\boldsymbol{\beta})}}{1 + e^{(\mathbf{x}_i'\boldsymbol{\beta})}} \\
\frac{d\log(p(\boldsymbol{\beta}, \mathbf{y}))}{d\boldsymbol{\beta}} &= \mathbf{X}'(\mathbf{y} - \mathbf{np})
\end{aligned}
$$

where $\mathbf{y} = (y_1, \ldots y_N)$, $\mathbf{n} = (n_1, \ldots n_N)$, and $\mathbf{p} = (p_1, \ldots p_N)$.

- Now we can move to R, implement HMC for this problem and compare it to the Metropolis-Hasting algorithm.