Advanced Simulation - Lecture 12

February 25th, 2020
Hamiltonian Monte Carlo

Hamiltonian Monte Carlo consists of the iteration of 2 steps,
1. Resample the momentum component $p$.
2. Propose a new position $\left( \begin{array}{c} q^* \\ p^* \end{array} \right) = \Psi \left( \begin{array}{c} q \\ p \end{array} \right) := N \left( \Psi^L \left( \begin{array}{c} q \\ p \end{array} \right) \right)$ by applying $L$ Leapfrog steps and flipping the momentum. Accept it with probability

$$\min \left[ 1, \exp \left( H(q, p) - H(q^*, p^*) \right) \right]$$

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

$\pi$ is invariant with respect to the Markov kernel proposed above. Moreover, $\pi$ is reversible with respect to the second step of the Markov kernel.

Proof

Since $q$ and $p$ are independent according to the target distribution $\pi$, it is clear that the first step keeps the target invariant. Let $P_2$ 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 $\pi$ is reversible with respect to $P_2$. 
Hamiltonian Monte Carlo

Proof (continued)

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

$$\int \int f(x, y) \pi(dx) K(x, dy) = \int \int f(x, y) \pi(dy) K(y, dx). \quad (1)$$

$P_2(x, dy)$ is non-zero only for $y = \Psi(x)$ and $y = x$, so

$$\int \int f(x, y) \pi(dx) P_2(x, dy)$$

$$= \int \int f(x, \Psi(x)) \min[1, e^{H(x) - H(\Psi(x))}] \pi(dx)$$

$$+ \int \int f(x, x) \left(1 - \min[1, e^{H(x) - H(\Psi(x))}] \right) \pi(dx)$$
Hamiltonian Monte Carlo

Proof (continued)

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

\[
\pi(dy) = \pi(dx) \times |\text{det}(\nabla \Psi)| \frac{\exp(-H(y))}{\exp(-H(x))} = \pi(dx) \cdot e^{H(x) - H(y)},
\]

and the first part of the above sum can be written

\[
\begin{align*}
&\int \int f(x, \Psi(x)) \min[1, e^{H(x) - H(\Psi(x))}] \pi(dx) \\
= &\int \int f(\Psi(y), y) \min[1, e^{H(\Psi(y)) - H(y)}] \pi(dx) \\
= &\int \int f(\Psi(y), y) \min[1, e^{H(\Psi(y)) - H(y)}] \cdot e^{H(y) - H(\Psi(y))} \pi(dy) \\
= &\int \int f(\Psi(y), y) \min[1, e^{H(y) - H(\Psi(y))}] \pi(dy).
\end{align*}
\]
The second part satisfies that
\[
\int \int f(x, x) \left(1 - \min[1, e^{H(x) - H(\Psi(x))}]\right) \pi(dx)
\]
\[
= \int \int f(y, y) \left(1 - \min[1, e^{H(y) - H(\Psi(y))}]\right) \pi(dy).
\]

By combining these two equations, we can see that the reversibility condition (1) 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 reversibility.
Hamiltonian Monte Carlo

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

**Example 1**

Consider a 2D Gaussian distribution with covariance matrix \( \Sigma = \begin{pmatrix} 1 & 0.98 \\ 0.98 & 1 \end{pmatrix} \). We let \( M = I_2 \), and do \( L = 25 \) leapfrog steps per iteration using stepsize \( \epsilon = 0.25 \). Figures 1 and 2 illustrate HMC, and compare it with random walk Metropolis.

![Figure: Position, momentum and the value of the Hamiltonian.](image)

Figure: Position, momentum and the value of the Hamiltonian.
Figure: Twenty iterations of the random walk Metropolis method (20 updates per iteration) and the HMC method (20 Leapfrog steps) for a highly correlated 2 dimensional Gaussian distribution. HMC is making much larger moves and mixes faster than random walk Metropolis.
Example 2

100 dimensional multivariate Gaussian distribution with independent components of standard deviations $0.01, ..., 1.00$. $\epsilon$ was randomly chosen at each iteration uniformly from $(0.0104, 0.0156)$, and $L=150$. HMC was compared with the random walk Metropolis. Figure 3 shows the last component.

Figure: The last component in a 100 dimensional Gaussian.
In the previous example, we have used randomly chosen step size $\varepsilon$, which will result in a random period length $L\varepsilon$.

This new Markov kernel is a mixture of Markov kernels for which $\pi$ is stationary. Therefore $\pi$ is also a stationary distribution with respect to this mixture Markov kernel.

By using a random period length instead of a fixed period length, we can avoid issues with periodicity that sometimes arise due to the deterministic nature of Hamiltonian dynamics.
Concentration in high dimensions

- To understand the reason for the better performance of HMC in high dimensions, we study the concentration of measure phenomenon.
- Suppose that $Z \sim N(0, I_d)$ is a $d$ dimensional standard normal random vector. Then the Euclidean norm of $Z$ satisfies that for every $t \geq 0$,

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

where $C$ is an absolute constant independent of $d$.
- This means that with high probability, $\|Z\| = \sqrt{d} + O(1)$, i.e. most of the probability is concentrated in a thin layer around the sphere of radius $\sqrt{d}$. 
Concentration in high dimensions

Figure: The density of $\pi$ for a 100 dimensional standard Gaussian, for first components $p_1, q_1$ (other components integrated out). The dashed line shows a possible Hamiltonian path in this case.
Concentration in high dimensions

- Note that here $H(z) = \frac{\|z\|^2}{2}$, and since $H(z)$ is preserved by the Hamiltonian, Hamiltonian dynamics moves around in circular arcs.

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

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

for some constant $C$ that depends on $\mu$ and $L$ but is independent of the dimension $d$.

- Hamiltonian is close to constant high probability density area, and HMC is very efficient in exploring this potentially complicated set automatically.
Hidden Markov Models - Outline

- Hidden Markov models, also called state space models.

- Various examples.

- Inference leads to high-dimensional integrals.
Time series

- Observations \((y_t)_{t \geq 1}\) assumed to be dependent, usually specified by an initial distribution: \(Y_1 \sim p(dy_1 | \theta)\), and a conditional distribution:

\[
Y_t | y_{1:t-1} \sim p(dy_t | y_{1:t-1}, \theta),
\]

where we use the notation \(y_{k:l} = (y_k, \ldots, y_l)\).

- The likelihood is given by

\[
\forall \theta \in \Theta \quad p(y_{1:t} | \theta) = p(y_1 | \theta) \prod_{k=2}^{t} p(y_k | y_{1:k-1}, \theta).
\]

- Put a prior on \(\theta\) and consider the problem of sampling from the posterior given \(y_{1:t}\).
Time series

- In simple cases, the likelihood can be computed point-wise.
- Example: Bayesian analysis of a Markov chain, in Chapter 3 of the lecture notes.
- ARCH(1) model: $y_1 \sim \mathcal{N}(0,1)$ and for all $t \geq 2$,

$$
y_t = \epsilon_t (h_t)^{1/2},$$

$$\epsilon_t \sim \mathcal{N}(0,1),$$

$$h_t = \alpha_0 + \alpha_1 y_{t-1}^2.$$

- In this case we can implement a Metropolis–Hastings algorithm to sample from $p(\theta \mid y_1:t)$, for each $t$.
- Or importance sampling to obtain estimates at each intermediate time $1 \leq s \leq t$. 
Hidden Markov Models

- We introduce \((X_t)_{t \geq 0}\) a latent/hidden/unobserved \(X\)-valued Markov process defined by its initial density

\[ X_0 | \theta \sim p(dx_0 | \theta), \]

and its homogeneous Markov transition kernel

\[ X_t | x_{t-1}, \theta \sim p(dx_t | x_{t-1}, \theta). \]

- Hence the law of the path/trajectory \(X_{1:t}\) is given by

\[
p(x_{0:t} | \theta) = p(x_0 | \theta) \prod_{k=1}^{t} p(x_k | x_{1:k-1}, \theta) \text{ (chain rule)}
\]

\[
= p(x_0 | \theta) \prod_{k=1}^{t} p(x_k | x_{k-1}, \theta) \text{ (Markov)}
\]
Hidden Markov Models

- Observations \( (Y_t)_{t \geq 1}, \mathbb{Y}-valued \), are assumed to be independent conditional upon \( (X_t)_{t \geq 0} \) and their conditional distributions satisfy

\[
Y_t \mid x_t, \theta \sim p(dy_t \mid x_t, \theta),
\]

i.e. the distribution of \( Y_t \) is independent of \( (X_k)_{k \neq t} \) conditional upon \( X_t = x_t \).

- Hence we have the law of observations given the hidden process,

\[
p(y_{1:t} \mid x_{1:t}, \theta) = \prod_{k=1}^{t} p(y_k \mid x_k, \theta) \text{ (cond. independent)}
\]
Hidden Markov Models

With a prior $p(\theta)$ on the parameters, the joint density is:

$$p(x_{0:T}, y_{1:T}, \theta) = p(\theta) p(x_0 | \theta) \prod_{t=1}^{T} p(x_t | x_{t-1}, \theta) \prod_{t=1}^{T} p(y_t | x_t, \theta).$$
Example: S&P 500 index

Figure: Daily returns $y_t = \log(p_t/p_{t-1})$ between 1984 and 1991.
Example: stochastic volatility model

- Latent stochastic volatility \((X_t)_{t \geq 1}\) of an asset is modeled through

\[
X_t = \phi X_{t-1} + \sigma V_t, \quad Y_t = \beta \exp(X_t) W_t
\]

where \(V_t, W_t \sim \mathcal{N}(0, 1)\).

- Popular alternative to ARCH and GARCH models (Engle, 2003 Nobel Prize).
Example: battery voltage

Figure: Current (input) and measured voltage (output) of a battery.
Example: phytoplankton — zooplankton

Figure: Filtering of the latent variables (top: P, bottom: Z).
Example: athletic records

Figure: Best two times of each year in women’s 3000m events.
Example: tracking
Example: tracking

- Markov model describing dynamic of the target

\[
\begin{pmatrix}
X_1^t \\
X_2^t
\end{pmatrix} =
\begin{pmatrix}
1 & \delta & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1 & \delta \\
0 & 0 & 0 & 1
\end{pmatrix}
\begin{pmatrix}
X_1^{t-1} \\
X_2^{t-1}
\end{pmatrix} + V_t, \quad V_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \Sigma_v),
\]

- Measurements provided by the radar

\[
Y_t = \tan^{-1}\left(\frac{X_1^t}{X_2^t}\right) + W_t, \quad W_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2).
\]
Specific Case: Finite State-Space HMM

- **Automatic speech recognition**: $Y_t$ is the speech signal, $X_t$ represents the word that is being spoken.
- **Activity recognition**: $Y_t$ represents a video frame, $X_t$ is the class of activity the person is engaged in (e.g., running, walking, sitting, etc.)
- **Part of speech tagging**: $Y_t$ represents a word, $X_t$ represents its part of speech (noun, verb, adjective, etc.)
- **Gene finding**: $Y_t$ represents the DNA nucleotides (A,C,G,T), $X_t$ represents whether we are inside a gene-coding region or not.

Specific algorithms allow to estimate $X_{0:t}$ given $y_{1:t}$ and to evaluate the likelihood of the parameters: Viterbi, forward-backward, Baum–Welch.
Specific Case: Linear Gaussian models

- Consider $X = \mathbb{R}^{d_x}$ and $Y = \mathbb{R}^{d_y}$. Let $X_t$ be defined by

$$X_t = AX_{t-1} + \varepsilon_t$$

for $\varepsilon \sim \mathcal{N}(0, \Sigma_x)$, and some matrices $A$ and $\Sigma_x$.

- Let the observations be defined by

$$Y_t = CX_t + \Sigma_y \eta_t$$

for $\eta \sim \mathcal{N}(0, \Sigma_y)$, and some matrices $C$ and $\Sigma_y$.

- Then the distribution of $X_{1:t}$ given $Y_{1:t}$ can be retrieved by “Kalman recursions”.

- The likelihood of the parameters $(A, C, \Sigma_x, \Sigma_y)$ can be evaluated exactly, which allows parameter estimation using standard techniques.
General case

Given $Y_{1:t} = y_{1:t}$ and $\theta$, inference on $X_{0:t}$ relies on

$$p(x_{0:t} \mid y_{1:t}, \theta) = \frac{p(x_{0:t}, y_{1:t} \mid \theta)}{p(y_{1:t} \mid \theta)} = \frac{p(x_{0:t} \mid \theta) p(y_{1:t} \mid x_{0:t}, \theta)}{p(y_{1:t} \mid \theta)}$$

with

$$p(x_{0:t} \mid \theta) = p(x_0 \mid \theta) \prod_{k=1}^{t} p(x_k \mid x_{k-1}, \theta)$$

$$p(y_{1:t} \mid x_{0:t}, \theta) = \prod_{k=1}^{t} p(y_k \mid x_k, \theta)$$

$$p(y_{1:t} \mid \theta) = \int_{X_t} p(x_{0:t}, y_{1:t} \mid \theta) \, dx_{0:t}.$$
General case

• **Proposition:** The posterior $p(x_{0:t} \mid y_{1:t}, \theta)$ satisfies

$$p(x_{0:t} \mid y_{1:t}, \theta) = p(x_{0:t-1} \mid y_{1:t-1}, \theta) \frac{p(x_t \mid x_{t-1}, \theta) p(y_t \mid x_t, \theta)}{p(y_t \mid y_{1:t-1}, \theta)}$$

where

$$p(y_t \mid y_{1:t-1}, \theta) = \int p(x_{0:t-1} \mid y_{1:t-1}, \theta) p(x_t \mid x_{t-1}, \theta) p(y_t \mid x_t, \theta) \, dx_{0:t}.$$ 

• **Proof.** Dropping the parameter $\theta$, we have

$$p(x_{0:t}, y_{1:t}) = p(x_{0:t-1}, y_{1:t-1}) p(x_t \mid x_{t-1}) p(y_t \mid x_t)$$

$$p(y_{1:t}) = p(y_{1:t-1}) p(y_t \mid y_{1:t-1})$$

so

$$p(x_{0:t} \mid y_{1:t}) = \frac{p(x_{0:t}, y_{1:t})}{p(y_{1:t})} = \frac{p(x_{0:t-1}, y_{1:t-1}) p(x_t \mid x_{t-1}) p(y_t \mid x_t)}{p(y_{1:t-1}) p(y_t \mid y_{1:t-1})}$$

and the expression for $p(y_t \mid y_{1:t-1})$ follows.
General case

- **Proposition:** The marginal posterior $p(x_t|y_{1:t})$ satisfies the following recursion

\[
p(x_t|y_{1:t-1}) = \int p(x_t|x_{t-1}) p(x_{t-1}|y_{1:t-1}) \, dx_{t-1}
\]

\[
p(x_t|y_{1:t}) = \frac{p(y_t|x_t) p(x_t|y_{1:t-1})}{p(y_t|y_{1:t-1})}
\]

where

\[
p(y_t|y_{1:t-1}) = \int p(y_t|x_t) p(x_t|y_{1:t-1}) \, dx_t.
\]

- This recursion can be implemented exactly for finite state-space HMM and linear Gaussian models.

- Otherwise these integrals are intractable.
General case

- In general, the filtering problem is thus intractable:

\[ \int \varphi(x_t) p(x_t | y_{1:t}, \theta) dx_t = \int \varphi(x_t) p(x_{1:t} | y_{1:t}, \theta) dx_{0:t} \]

\[ = \int \varphi(x_t) p(x_0 | \theta) \prod_{k=1}^{t} p(x_k | x_{k-1}, \theta) \prod_{k=1}^{t} p(y_k | x_k, \theta) dx_{0:t} \]

\[ p(y_{1:t} | \theta) \]

- The numerator is a \( t \times \text{dim}(X) \) dimensional integral.
- The denominator, the marginal likelihood is also intractable:

\[ p(y_{1:t} | \theta) = \int p(x_{0:t}, y_{1:t} | \theta) dx_{0:t} \]

\[ = \int p(x_0 | \theta) \prod_{k=1}^{t} p(x_k | x_{k-1}, \theta) \prod_{k=1}^{t} p(y_k | x_k, \theta) dx_{0:t} \]

- Thus we cannot compute it pointwise, e.g. to perform Metropolis–Hastings algorithm on the parameter space.
General case

- The historical approach consists in performing Gibbs sampling on the joint space of $\theta$ and $X_{0:t}$.

- Alternate between sampling from $\theta | x_{0:t}, y_{1:t}$, with conditional distribution

$$p(\theta | x_{0:t}, y_{1:t}) \propto p(\theta) p(x_{0:t}, y_{1:t} | \theta)$$

$$= p(\theta) p(x_0 | \theta) \prod_{k=1}^{t} p(x_k | x_{k-1}, \theta) \prod_{k=1}^{t} p(y_k | x_k, \theta)$$

which can (perhaps) be evaluated pointwise.

- And sampling from $p(x_{0:t} | y_{1:t}, \theta)$. How?
General case

- Sampling from \( p(x_{0:t} \mid y_{1:t}, \theta) \) can be done by iteratively sampling \( x_k \) given \( x_{k-1}, y_k, x_{k+1} \) and \( \theta \).

- Indeed

\[
p(x_k \mid x_{-k}, y_{1:t}, \theta) = p(x_k \mid x_{k-1}, y_k, x_{k+1}, \theta) \propto p(x_k \mid x_{k-1}, \theta) p(y_k, x_{k+1} \mid x_k, \theta)
= p(x_k \mid x_{k-1}, \theta) p(x_{k+1} \mid x_k, \theta) p(y_k \mid x_k, \theta)
\]

and (perhaps) we can evaluate this density point-wise.

- In which case, we could use Metropolis–Hastings to update each component of \( X_{0:t} \) given the others.

- By definition, the components of \( X_{0:t} \) are highly correlated, thus this Gibbs sampling approach will fail (remember the bivariate normal!!).
Objects of practical interest

Various by-products of the joint posterior $p(\theta, x_{0:t} | y_{1:t})$:

- prediction under parameter uncertainty through $p(x_{t+1} | y_{1:t})$,

\[
p(x_{t+1} | y_{1:t}) = \int_{\Theta} \int_{\mathcal{X}_{t+1}} p(x_{t+1} | x_t, \theta) \, p(d\theta, dx_{0:t} | y_{1:t}).
\]

- predictive checking through $\mathbb{P}(Y_{t+1} \leq y_{t+1} | y_{1:t})$,

\[
\mathbb{P}(Y_{t+1} < y_{t+1} | y_{1:t}) = \int_{\Theta} \int_{\mathcal{X}} \int_{-\infty}^{y_{t+1}} p(dy | x_{t+1}, \theta) \, p(dx_{t+1}, d\theta | y_{1:t}).
\]

- sequential model comparison through $p(y_{1:t})$.

\[
p(y_{1:t}) = \int_{\Theta} \int_{\mathcal{X}_{t+1}} p(d\theta, dx_{0:t}, y_{1:t})
\]
Summary and What’s Next

- Usually, batch estimation of $p(\theta, x_{1:T} \mid y_{1:T})$ using MCMC performs poorly because of high correlations between components.

- Filtering given a fixed $\theta$, i.e. approximating $p(x_{1:t} \mid y_{1:t}, \theta)$, can be efficiently performed using particle filters.

- We’ll see that particle filters also provide estimators of the likelihood, which can be used to estimate $\theta$.

- Particle Markov chain Monte Carlo for batch estimation of $p(\theta, x_{1:T} \mid y_{1:T})$ (Andrieu, Doucet, Holenstein, 2010).