Hamiltonian Monte Carlo methods

A Riemannian geometry perspective

Emile Mathieu, Kimia Nadjahi
April 11, 2017

Ecole des Ponts ParisTech
Problem Statement
Problem Statement

**Intractable density:**
\[ p(\beta) = \tilde{p}(\beta) / \int \tilde{p}(\beta) d\beta \]

**Metropolis-Hastings:**

- Define an ergodic Markov process with stationary distribution \( p(\beta) \).
- Transitions \( \beta \rightarrow \beta^* \) proposed with density \( q(\beta^* | \beta) \) accepted with probability

\[
\alpha(\beta, \beta^*) = \min \left\{ 1, \frac{\tilde{p}(\beta^*) q(\beta | \beta^*)}{\tilde{p}(\beta) q(\beta^* | \beta)} \right\}
\]

**What proposal distribution \( q \) ?**

Typically, random walk: \( q(\beta^* | \beta) = \mathcal{N}(\beta^* | \beta, \Lambda) \)

<table>
<thead>
<tr>
<th>Low ( |\Lambda| )</th>
<th>High ( |\Lambda| )</th>
</tr>
</thead>
<tbody>
<tr>
<td>High acceptance rate</td>
<td>Low acceptance rate</td>
</tr>
<tr>
<td>Highly correlated samples</td>
<td>Not so many correlated samples</td>
</tr>
</tbody>
</table>
Hamiltonian Monte Carlo
Hamiltonian Monte Carlo (I)

**Goal:**
Make large transitions accepted with high probability.

**How to:**
- Independent auxiliary variable: \( p \sim \mathcal{N}(p|0, M) \)
- Joint density: \( p(\beta, p) = p(\beta)p(p) = p(\beta)\mathcal{N}(p|0, M) \)
- The negative joint log-probability is (with \( \mathcal{L}(\beta) = \log\{p(\beta)\} \))

\[
H(\beta, p) = -\mathcal{L}(\beta) + \frac{1}{2} \log\{(2\pi)^D|M|\} + \frac{1}{2} p^T M^{-1} p
\]

(1)

where \( \mathcal{L}(\beta) = \log\{p(\beta)\} \)
Hamiltonian Monte Carlo (II)

- Hamilton’s equations:

\[
\frac{d\beta}{d\tau} = \frac{\partial H}{\partial p} = M^{-1} p, \quad \frac{dp}{d\tau} = -\frac{\partial H}{\partial \beta} = \nabla_{\beta} L(\beta)
\]

(2)

- Numerical integrator (Stormer-Verlet or leapfrog):

\[
p^{\tau+\frac{\epsilon}{2}} = p^\tau + \frac{\epsilon}{2} \nabla_{\beta} L\{\beta^\tau\}
\]

(3)

\[
\beta^{\tau+\epsilon} = \beta^\tau + \epsilon M^{-1} p^{\tau+\frac{\epsilon}{2}}
\]

(4)

\[
p^{\tau+\epsilon} = p^{\tau+\frac{\epsilon}{2}} + \frac{\epsilon}{2} \nabla_{\beta} L\{\beta^{\tau+\epsilon}\}
\]

(5)

- Acceptance probability: \(\min(1, \exp\{-H(\beta^*, p^*) + H(\beta, p)\})\)

Hyperparameters:

- Step size \(\epsilon\) and number of integration steps: via acceptance rate
- Yet, choice of mass matrix \(M\) is critical.
Geometric concepts
Geometric concepts

**Goal:**
Automatically determine $M$ at each step.

**Fisher-Rao metric:**
Distance between parametrized density functions

$$
\text{KL}(p(y; \beta)||p(y; \beta + \delta \beta)) \simeq \delta \beta^T G(\beta) \delta \beta
$$

where $G(\beta) = -\mathbb{E}_{y|\beta} \left[ \frac{\partial^2}{\partial \beta^2} \log \{ p(y|\beta) \} \right]$

Fisher information matrix $G(\beta)$ is p.d. metric defining a Riemann manifold.

**General metric tensor**
With a Bayesian perspective:

$$
G(\beta) = -\mathbb{E}_{y|\beta} \left[ \frac{\partial^2}{\partial \beta^2} \log \{ p(y, \beta) \} \right]
$$
Riemann manifold Hamiltonian Monte Carlo
The Hamiltonian is

$$H(\beta, p) = -\mathcal{L}(\beta) + \frac{1}{2} \log \{(2\pi)^D |G(\beta)|\} + \frac{1}{2} p^T G(\beta)^{-1} p \quad (6)$$

Marginal density:

$$p(\beta) \propto \int \exp \{-H(\beta, p)\} \, dp = \exp \{\mathcal{L}(\beta)\}$$

Joint density: $p(\beta, p) = p(\beta)p(p|\beta) = p(\beta)\mathcal{N}(p|0, G(\beta))$

Hamilton’s equations:

$$\frac{d\beta_i}{d\tau} = \frac{\partial H}{\partial p_i} = \{G(\beta)^{-1} p\}_i \quad (7)$$

$$\frac{dp_i}{d\tau} = -\frac{\partial H}{\partial \beta_i} = \frac{\partial \mathcal{L}(\beta)}{\partial \beta_i} - \frac{1}{2} \text{tr} \left\{ G(\beta)^{-1} \frac{\partial G(\beta)}{\partial \beta_i} \right\}$$

$$+ \frac{1}{2} p^T G(\beta)^{-1} \frac{\partial G(\beta)}{\partial \beta_i} G(\beta)^{-1} p \quad (8)$$
Numerical integrator (generalized leapfrog):

\[ p^{\tau + \frac{\epsilon}{2}} = p^\tau - \frac{\epsilon}{2} \nabla_\beta H \left\{ \beta^\tau, p^{\tau + \epsilon/2} \right\} \]  \hspace{1cm} (9)

\[ \beta^{\tau + \epsilon} = \beta^\tau + \frac{\epsilon}{2} \left[ \nabla_p H \left\{ \beta^\tau, p^{\tau + \epsilon/2} \right\} + \nabla_p H \left\{ \beta^{\tau + \epsilon}, p^{\tau + \epsilon/2} \right\} \right] \]  \hspace{1cm} (10)

\[ p^{\tau + \epsilon} = p^{\tau + \epsilon/2} - \frac{\epsilon}{2} \nabla_\beta H \left\{ \beta^{\tau + \epsilon}, p^{\tau + \epsilon/2} \right\} \]  \hspace{1cm} (11)

Acceptance probability:

\[ \min(1, \exp\{-H(\beta^*, p^*) + H(\beta, p)\}) \]
Application: Bayesian logistic regression
Bayesian logistic regression (BLR)

Probabilistic model

- Likelihood:
  \[ p(y = 1|X) = \eta(X^T\beta) \]
  with \( \eta : x \mapsto (1 + e^{-x})^{-1} \)

- Prior: \( \beta \sim \mathcal{N}(0, \alpha I) \)

- Metric tensor:
  \[ G(\beta) = X^T \Lambda X + \alpha^{-1} I \]
  with \( \Lambda_{i,i} = \eta(\beta^T X_{n,.}) \{1 - \eta(\beta^T X_{n,.})\} \)

Figure 1: DAG for the Bayesian logistic regression.
Iterative Weighted Least Squares (IWLS)

- MLE of BLR obtained with Newton-Raphson method:
  \[
  \hat{\beta}^{(t)} = (((\alpha I)^{-1} + X^T \Lambda X)^{-1}(X^T \Lambda \tilde{y}(\hat{\beta}^{(t-1)}))
  
  \hat{\Sigma}_{\beta}^{(t)} = G(\beta)^{-1} = (((\alpha I)^{-1} + X^T \Lambda X)^{-1}
  
  where \( \tilde{y}(\hat{\beta}^{(t-1)}) = X\hat{\beta}^{(t-1)} + \Lambda^{-1}(y - \eta(\beta^T X)) \)

- Combination of the MCMC and IWLS iteration schemes:
  1. Initialization: \( \beta = \beta^{(0)}, \ t = 1 \)
  2. Sample \( \beta_{\text{new}} \) from \( \mathcal{N}(\hat{\beta}^{(t)}, \hat{\Sigma}_{\beta}^{(t)}) \)
  3. Accept it with probability \( \alpha(\beta^{(t-1)}, \beta_{\text{new}}) \) and set \( \beta^{(t)} = \beta_{\text{new}} \); otherwise, \( \beta^{(t)} = \beta^{(t-1)} \)
  4. Do \( t := t + 1 \) and return to Step 2
Representation of BLR with auxiliary variables [Helmes and Hold, 2005]:

\[ y_i = \text{sgn}(z_i) \]
\[ z_i = X_i \beta + \epsilon_i \]
\[ \epsilon_i \sim \mathcal{N}(0, \lambda_i) \]
\[ \lambda_i = (2\psi_i)^2 \]
\[ \psi_i \sim \text{KS} \]
\[ \beta \sim \mathcal{N}(0, \alpha I) \]

\( \beta|z, \lambda \sim \mathcal{N}(B, V) \) with \( B, V \) as in WLS, \( z_i|\beta, X_i, y_i, \lambda_i \sim \) truncated normal, \( \lambda_i|z_i, \beta \) sampled with rejection sampling

Block Gibbs sampler:

1. Update \( \{z, \beta\} \) jointly given \( \lambda \)
2. Update \( \lambda|z, \beta \)
Component-wise random symmetric walk
At iteration $i$ for component $k$:

1. Sample $\tilde{\beta}_{i+1}^k \sim \mathcal{N}(\beta_i^k, \sigma_k)$
2. Set $\beta_{i+1}^k = \tilde{\beta}_{i+1}^k$ with probability $\tilde{p}(\tilde{\beta}_{i+1}) / \tilde{p}(\beta_i)$
3. Otherwise $\beta_{i+1}^k = \beta_i^k$

Component-wise adaptive variance
Every 100 samples after burn-in:

- If $\text{currentAcceptanceRate}^k > \gamma_{\text{max}}$
  $\sigma_k = 1.2 \times \sigma_k$
- Else if $\text{currentAcceptanceRate}^k < \gamma_{\text{min}}$
  $\sigma_k = 0.8 \times \sigma_k$
Effective Sample Size (ESS)

\[ ESS = \frac{N}{1 + 2 \sum_k \gamma(k)} \]

where \( N \) number of samples after burn-in, \( \gamma(k) \) autocorrelation of lag \( k \)

Ideally, \( ESS = N \).

Trade-off time and quality: ratio time / min(ESS)
Averaged results: sampling experiments repeated 10 times

<table>
<thead>
<tr>
<th>Method</th>
<th>Time</th>
<th>$ESS$ (min, median, max)</th>
<th>time / min($ESS$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metropolis</td>
<td>10.4</td>
<td>(305.5, 653.6, 801)</td>
<td>0.034</td>
</tr>
<tr>
<td>Auxiliary variables</td>
<td>589.4</td>
<td>(710.6, 1199.4, 1715.2)</td>
<td>0.83</td>
</tr>
<tr>
<td>HMC</td>
<td>44.9</td>
<td>(3349, 3634, 4141)</td>
<td>0.0134</td>
</tr>
<tr>
<td>IWLS</td>
<td>17.9</td>
<td>(21.78, 96.34, 322.4)</td>
<td>0.83</td>
</tr>
<tr>
<td>RMHMC</td>
<td>164.1</td>
<td>(4865, 5000, 5000)</td>
<td>0.034</td>
</tr>
</tbody>
</table>

**Table 1:** Heart data set – comparison of sampling methods
Results

Figure 2: Autocorrelation plots: 1st covariate, 1000 samples of Heart data

(a) Adaptive MH  (b) Auxiliary Gibbs  (c) IWLS
(d) HMC  (e) RMHMC
Conclusion
Thank you for your attention!

Questions?