Bayesian Statistics I

Lecture 9 - HMC and Variational Inference

Mattias Villani

Department of Statistics
Stockholm University

Department of Computer and Information Science
Linköping University

mattiasvillani.com @matvil mattiasvillani

Lecture overview

- Hamiltonian Monte Carlo
- Variational Inference

Hamiltonian Monte Carlo

- When $\theta = (\theta_1, \ldots, \theta_p)$ is **high-dimensional**, $p(\theta|\mathrm{y})$ usually located in some subregion of $\mathbb{R}^p$ with complicated geometry.
- MH: hard to find good proposal distribution $q\left(\cdot|\theta^{(i-1)}\right)$.
- MH: use very small step sizes otherwise too many rejections.
- **Hamiltonian Monte Carlo (HMC)**:
  - distant proposals **and**
  - high acceptance probabilities.
- HMC: add extra **momentum** parameters $\phi = (\phi_1, \ldots, \phi_p)$ and sample from

$$p(\theta, \phi|\mathrm{y}) = p(\theta|\mathrm{y})\, p(\phi)$$

.

Hamiltonian Monte Carlo

- Physics: **Hamiltonian** system $H(\theta, \phi) = U(\theta) + K(\phi)$, where U is the **potential energy** and K is the **kinetic energy**.
- **Hamiltonian Dynamics**

$$\frac{d\theta_i}{dt} = \frac{\partial H}{\partial \phi_i} = \frac{\partial K}{\partial \phi_i},$$
$$\frac{d\phi_i}{dt} = -\frac{\partial H}{\partial \theta_i} = -\frac{\partial U}{\partial \theta_i}$$

- Hockey puck sliding over a friction-less surface: illustration.
- Use $U(\theta) = -\log\left[p(\theta)\, p(\mathrm{y}|\theta)\right]$.
- Use $\phi \sim N(0, \mathrm{M})$ where M is the mass matrix and

$$K(\phi) = -\log\left[p(\phi)\right] = \frac{1}{2}\phi^T \mathrm{M}^{-1}\phi + \mathrm{const}$$

- If we could propose θ in continuous time (spoiler: we can't), the acceptance probability would be one.

Hamiltonian Monte Carlo

- **Hamiltonian Dynamics**

$$\frac{d\theta_i}{dt} = \left[\mathrm{M}^{-1}\phi\right]_i,$$
$$\frac{d\phi_i}{dt} = \frac{\partial \log p\left(\theta|\mathrm{y}\right)}{\partial\theta_i}$$

which can be simulated using the **leapfrog algorithm**

$$\phi_i\left(t+\frac{\varepsilon}{2}\right) = \phi_i\left(t\right) + \frac{\varepsilon}{2}\frac{\partial \log p\left(\theta(t)|\mathrm{y}\right)}{\partial\theta_i},$$
$$\theta\left(t+\varepsilon\right) = \theta\left(t\right) + \varepsilon\mathrm{M}^{-1}\phi(t),$$
$$\phi_i\left(t+\varepsilon\right) = \phi_i\left(t+\frac{\varepsilon}{2}\right) + \frac{\varepsilon}{2}\frac{\partial \log p\left(\theta(t)|\mathrm{y}\right)}{\partial\theta_i},$$

where ε is the step size.

- **Discretization** $\Rightarrow$ acceptance probability drops with ε.

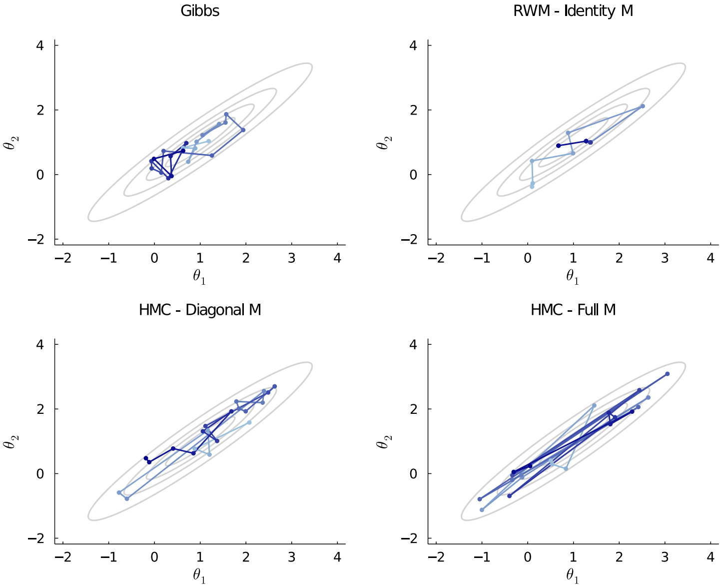
The Hamiltonian Monte Carlo algorithm

- Initialize $\theta^{(0)}$ and iterate for $i = 1, 2, \ldots$

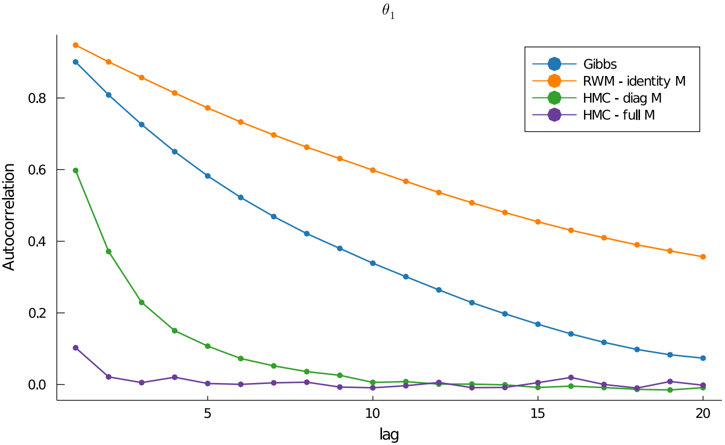
  1. Sample the starting **momentum** $\phi_s \sim N(0, \mathrm{M})$
  2. Simulate new values for (θ_p, ϕ_p) by iterating the **leapfrog algorithm** L times, starting in $\left(\theta^{(i-1)}, \phi_s\right)$.
  3. Compute the **acceptance probability**
  $$\alpha = \min\left(1, \frac{p(\mathrm{y}|\theta_p)p(\theta_p)}{p(\mathrm{y}|\theta^{(i-1)})p(\theta^{(i-1)})}\frac{p(\phi_p)}{p(\phi_s)}\right)$$
  4. With probability α set $\theta^{(i)} = \theta_p$ and $\theta^{(i)} = \theta^{(i-1)}$ otherwise.

- **Tuning parameters**: 1. **stepsize** ε, 2. **number of leapfrog** iterations L and 3. **mass matrix** M. **No U-turn**.

Comparing algorithms for bivariate normal



Comparing algorithms for bivariate normal



Variational Inference

- Let $\theta = (\theta_1, ..., \theta_p)$. Approximate the posterior $p(\theta|y)$ with a (simpler) distribution $q(\theta)$.
- Before: **Normal approximation** from optimization: $q(\theta) = N\left[\tilde{\theta}, J_{\mathrm{y}}^{-1}(\tilde{\theta})\right]$.
- **Mean field Variational Inference (VI)**: $q(\theta) = \prod_{i=1}^{p} q_i(\theta_i)$
- **Parametric VI**: Parametric family $q_\lambda(\theta)$ with parameters λ
- Find the $q(\theta)$ that **minimizes the Kullback-Leibler distance** between the true posterior p and the approximation q:

$$KL(q,p) = \int q(\theta) \ln \frac{q(\theta)}{p(\theta|y)} d\theta = E_q\left[\ln \frac{q(\theta)}{p(\theta|y)}\right].$$

Mean field approximation

- **Mean field VI** is based on factorized approximation:

$$q(\theta) = \prod_{j=1}^{p} q_j(\theta_j)$$

- **No specific functional forms** are assumed for the $q_j(\theta)$.

- **Optimal densities** can be shown to satisfy:

$$q_j(\theta_j) \propto \exp\left(E_{-\theta_j} \ln p(\mathrm{y}, \theta)\right)$$

where $E_{-\theta_j}(\cdot)$ is the expectation with respect to $\prod_{k \neq j} q_k(\theta_k)$.

- **Structured mean field approximation**. Group subset of parameters in tractable blocks. Similar to Gibbs sampling.

Mean field approximation - algorithm

- Initialize: $q_2^*(\theta_2), \ldots, q_M^*(\theta_p)$
- Repeat until convergence:
  - $q_1^*(\theta_1) \leftarrow \frac{\exp\left[E_{-\theta_1} \ln p(\mathrm{y},\theta)\right]}{\int \exp\left[E_{-\theta_1} \ln p(\mathrm{y},\theta)\right] d\theta_1}$
  - $\vdots$
  - $q_p^*(\theta_p) \leftarrow \frac{\exp\left[E_{-\theta_p} \ln p(\mathrm{y},\theta)\right]}{\int \exp\left[E_{-\theta_p} \ln p(\mathrm{y},\theta)\right] d\theta_p}$
- Note: no assumptions about parametric form of the $q_i(\theta)$.
- Optimal $q_i(\theta)$ often **turn out** to be parametric (normal etc).
- Just update hyperparameters in the optimal densities.

Mean field approximation - Normal model

- **Model**: $X_i|\theta,\sigma^2 \overset{\text{iid}}{\sim} N(\theta,\sigma^2)$.
- **Prior**: $\theta \sim N(\mu_0,\tau_0^2)$ **independent** of $\sigma^2 \sim \text{Inv}-\chi^2(\nu_0,\sigma_0^2)$.
- **Mean-field approximation**: $q(\theta,\sigma^2) = q_\theta(\theta)\cdot q_{\sigma^2}(\sigma^2)$.
- Optimal densities

$$q_\theta^*(\theta) \propto \exp\left[E_{q(\sigma^2)} \ln p(\theta,\sigma^2,\mathrm{x})\right]$$
$$q_{\sigma^2}^*(\sigma^2) \propto \exp\left[E_{q(\theta)} \ln p(\theta,\sigma^2,\mathrm{x})\right]$$

Normal model - VB algorithm

- **Variational density for σ^2**

$$\sigma^2 \sim Inv-\chi^2\left(\tilde{\nu}_n, \tilde{\sigma}_n^2\right)$$

where $\tilde{\nu}_n = \nu_0 + n$ and $\tilde{\sigma}_n^2 = \frac{\nu_0\sigma_0^2 + \sum_{i=1}^n (x_i - \tilde{\mu}_n)^2 + n\cdot\tilde{\tau}_n^2}{\nu_0 + n}$

- **Variational density for θ**

$$\theta \sim N\left(\tilde{\mu}_n, \tilde{\tau}_n^2\right)$$

where

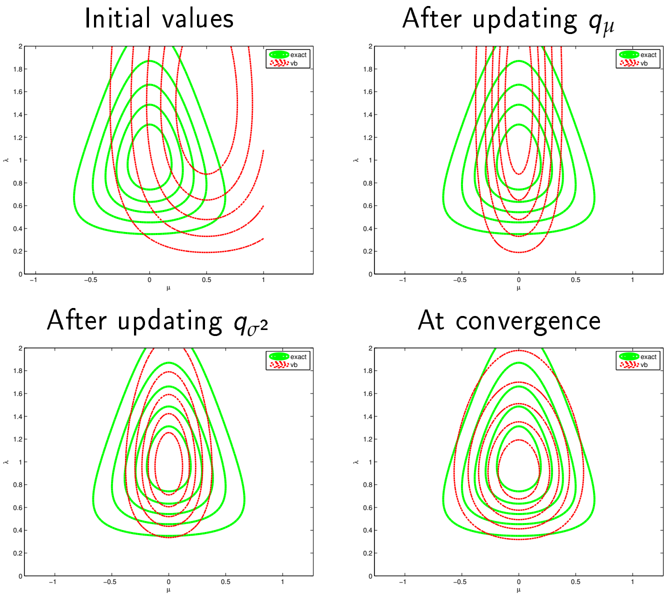
$$\tilde{\tau}_n^2 = \frac{1}{\frac{n}{\tilde{\sigma}_n^2} + \frac{1}{\tau_0^2}}$$

$$\tilde{\mu}_n = \tilde{w}\bar{x} + (1-\tilde{w})\mu_0,$$

where

$$\tilde{w} = \frac{\frac{n}{\tilde{\sigma}_n^2}}{\frac{n}{\tilde{\sigma}_n^2} + \frac{1}{\tau_0^2}}$$

Normal example from Murphy ($\lambda = 1/\sigma^2$)



Probit regression

- **Model**:

$$\Pr\left(y_i = 1|\mathrm{x}_i\right) = \Phi(\mathrm{x}_i^T\beta)$$

- **Prior**: $\beta \sim N(0, \Sigma_\beta)$. For example: $\Sigma_\beta = \tau^2 I$.
- **Latent variable formulation** with $u = (u_1, ..., u_n)'$

$$\mathrm{u}|\beta \sim N(\mathrm{X}\beta, 1)$$

and

$$y_i = \begin{cases} 0 & \text{if } u_i \leq 0 \\ 1 & \text{if } u_i > 0 \end{cases}$$

- Factorized **variational approximation**

$$q(\mathrm{u}, \beta) = q_{\mathrm{u}}(\mathrm{u})q_\beta(\beta)$$

VI for probit regression

- **VI posterior**

$$\beta \sim N\left(\tilde{\mu}_\beta, \left(\mathsf{X}^T\mathsf{X} + \Sigma_\beta^{-1}\right)^{-1}\right)$$

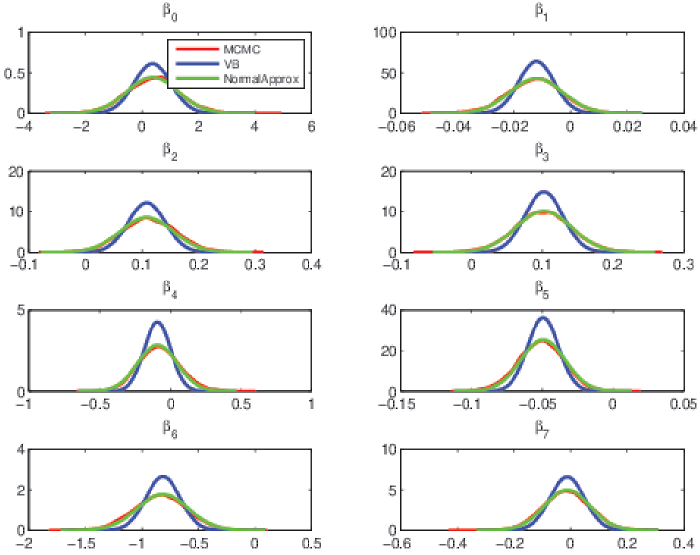
where

$$\tilde{\mu}_\beta = \left(\mathsf{X}^T\mathsf{X} + \Sigma_\beta^{-1}\right)^{-1}\mathsf{X}^T\tilde{\mu}_{\mathsf{u}}$$

and

$$\tilde{\mu}_{\mathsf{u}} = \mathsf{X}\tilde{\mu}_\beta + \frac{\phi\left(\mathsf{X}\tilde{\mu}_\beta\right)}{\Phi\left(\mathsf{X}\tilde{\mu}_\beta\right)^{\mathsf{y}}\left[\Phi\left(\mathsf{X}\tilde{\mu}_\beta\right) - 1_n\right]^{1_n - \mathsf{y}}}.$$

Probit example (n=200 observations)



Probit example