Bayesian Learning

Lecture 9 - Hamiltonian Monte Carlo and Variational Inference

Mattias Villani

Department of Statistics
Stockholm University

mattiasvillani.com @matvil @matvil mattiasvillani

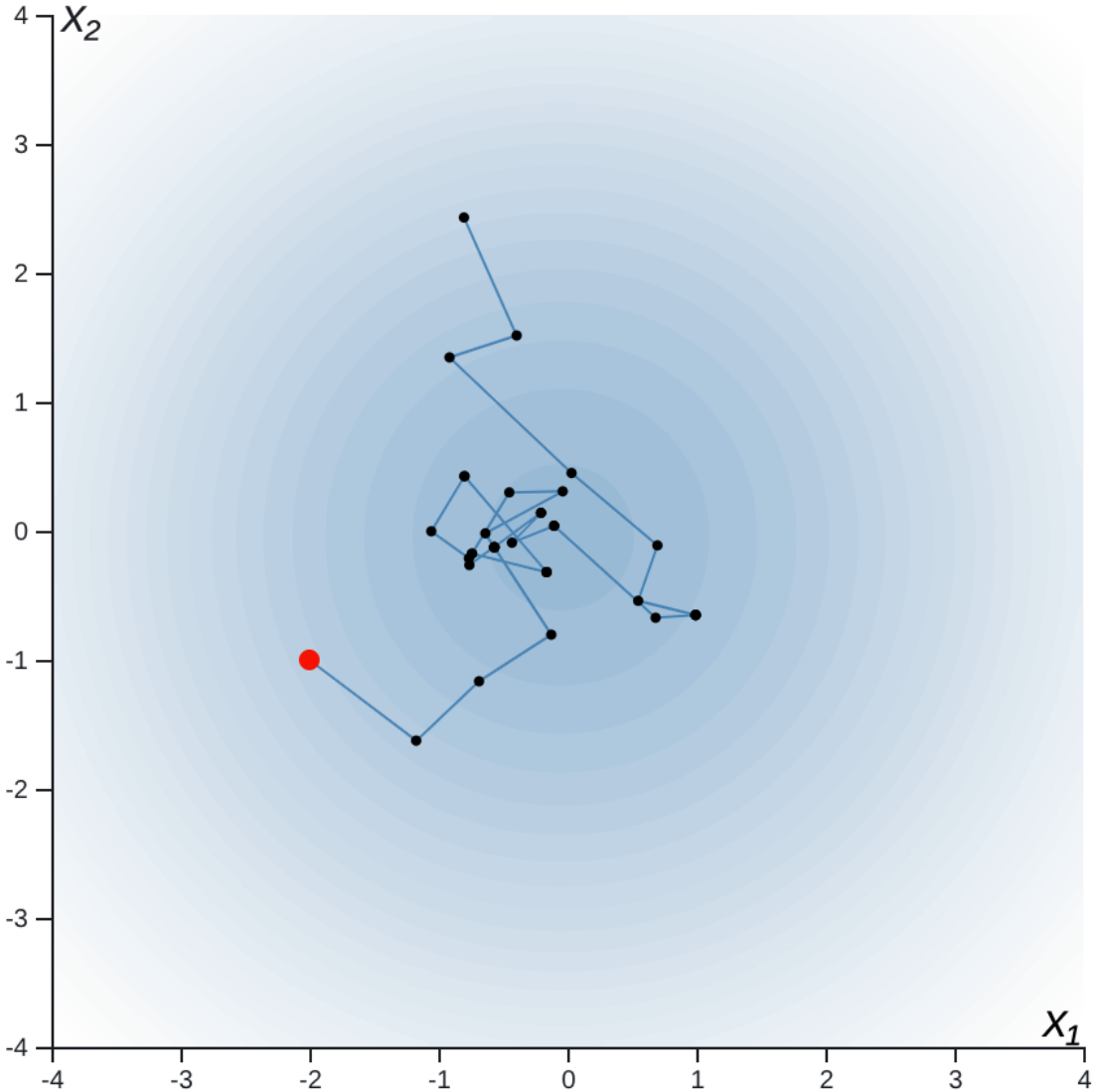
Lecture overview

- Hamiltonian Monte Carlo
- Variational Inference

Hamiltonian Monte Carlo

- When $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top$ is **high-dimensional**, $p(\boldsymbol{\theta}|\mathbf{y})$ usually located in some subregion of $\mathbb{R}^p$ with complicated geometry.
- **Metropolis-Hastings** (**MH**):
  - hard to find good proposal distribution $q\left(\cdot|\boldsymbol{\theta}^{(i-1)}\right)$.
  - RWM: must use small step size to avoid too many rejections.
- **Hamiltonian Monte Carlo** (**HMC**):
  - distant proposals **and**
  - high acceptance probabilities.

Interactive - Random Walk Metropolis



HMC is based on ideas from physics

- Image a ball sliding over a friction-less surface: illustration.
- **Location** of an object $\boldsymbol{\theta}$ $(p \times 1)$ (2D coordinates when $p = 2$).
- **Momentum** $\boldsymbol{\phi}$ $(p \times 1)$ - (mass $\times$ directional speed)
- **Hamiltonian** system
  - **Total energy** $H(\boldsymbol{\theta}, \boldsymbol{\phi}) = U(\boldsymbol{\theta}) + K(\boldsymbol{\phi})$
  - U is the **potential energy** ("stored" energy, given by position)
  - K is the **kinetic energy** (energy from movement)
- **Hamiltonian Dynamics**: how the position and momentum **changes over time**:

$$\frac{d\boldsymbol{\theta}}{dt} = \frac{\partial H}{\partial \boldsymbol{\phi}}$$
$$\frac{d\boldsymbol{\phi}}{dt} = -\frac{\partial H}{\partial \boldsymbol{\theta}}$$

Hamiltonian Monte Carlo

- HMC: MH algorithm to sample from posterior

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

  using **Hamiltonian dynamics to generate proposal $\theta^\star$**.

- Trick: add momentum parameters $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)^\top$ and sample from

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

- **Potential energy** is negative log posterior

$$U(\boldsymbol{\theta}) = -\log p(\boldsymbol{\theta}|\mathbf{y})$$

- **Momentum**: $\boldsymbol{\phi} \sim N(\mathbf{0}, \mathbf{M})$ where $\mathbf{M}$ is the mass matrix and

$$K(\boldsymbol{\phi}) = -\log p(\boldsymbol{\phi})$$

- If we could propose $\boldsymbol{\theta}$ in continuous time (spoiler: we can't), the MH acceptance probability would be one.

Approximate Dynamics via the Leapfrog algorithm

- **Hamiltonian Dynamics**

$$\frac{d\boldsymbol{\theta}}{dt} = \mathbf{M}^{-1}\boldsymbol{\phi}$$
$$\frac{d\boldsymbol{\phi}}{dt} = \frac{\partial \log p(\boldsymbol{\theta}|\mathbf{y})}{\partial \boldsymbol{\theta}}$$

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

Leapfrog algorithm

Function LEAPFROG($\boldsymbol{\theta}^{(0)}, \boldsymbol{\phi}^{(0)}, L, \epsilon, \mathbf{M}$)

for l in $1{:}L$ **do**

$$\tilde{\boldsymbol{\phi}}^{(l)} = \boldsymbol{\phi}^{(l-1)} + \frac{\epsilon}{2} \frac{\partial \log p\left(\mathbf{y}|\boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} \Big|_{\boldsymbol{\theta}^{(l-1)}}$$

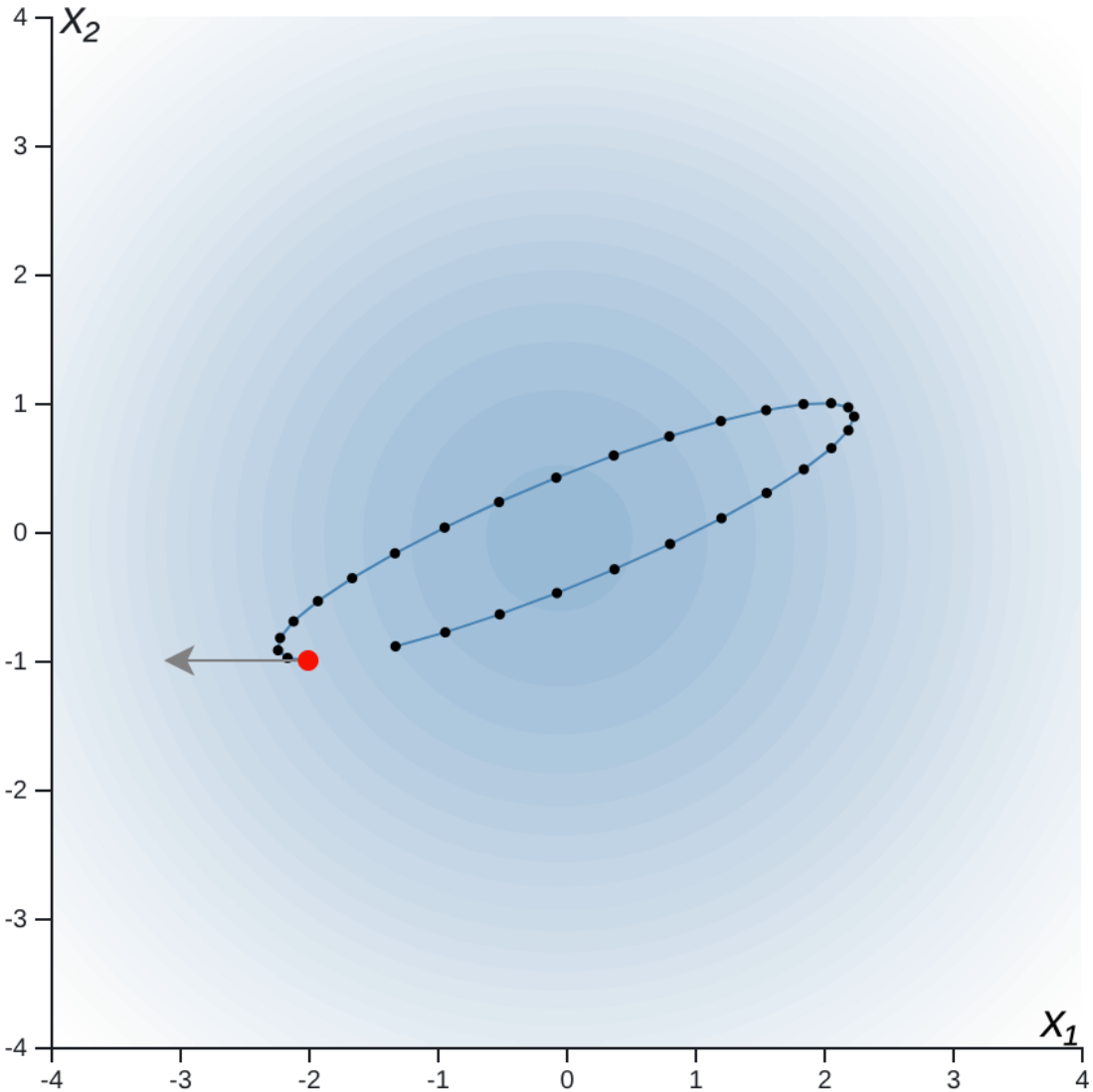
$$\boldsymbol{\theta}^{(l)} = \boldsymbol{\theta}^{(l-1)} + \epsilon \cdot \mathbf{M}^{-1} \tilde{\boldsymbol{\phi}}^{(l)}$$

$$\boldsymbol{\phi}^{(l)} = \tilde{\boldsymbol{\phi}}^{(l)} + \frac{\epsilon}{2} \frac{\partial \log p\left(\mathbf{y}|\boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}\right)}{\partial \theta_i} \Big|_{\boldsymbol{\theta}^{(l)}}$$

end

Output: final position and momentum $\boldsymbol{\theta}^{(L)}, \boldsymbol{\phi}^{(L)}$.

Interactive - Leapfrog



The Hamiltonian Monte Carlo algorithm

for i in $1{:}(m+b)$ **do**

Draw initial momentum $\boldsymbol{\phi}_s \sim N(\mathbf{0}, \mathbf{M})$

Compute proposal with leaf-frog algorithm

$\boldsymbol{\theta}^\star, \boldsymbol{\phi}^\star = \text{LeapFrog}(\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\phi}_s, L, \epsilon, \mathbf{M})$

Accept/reject proposal

Compute the acceptance probability

$$\alpha \leftarrow \min\left(1, \frac{p(\mathbf{y}|\boldsymbol{\theta}^\star)p(\boldsymbol{\theta}^\star)}{p(\mathbf{y}|\boldsymbol{\theta}^{(i-1)})p(\boldsymbol{\theta}^{(i-1)})}\frac{p(\boldsymbol{\phi}^\star)}{p(\boldsymbol{\phi}_s)}\right)$$

Draw $u \sim \text{Uniform}(0,1)$

if $u < \alpha$ **then**

$\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^\star$

else

$\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^{(i-1)}$

end

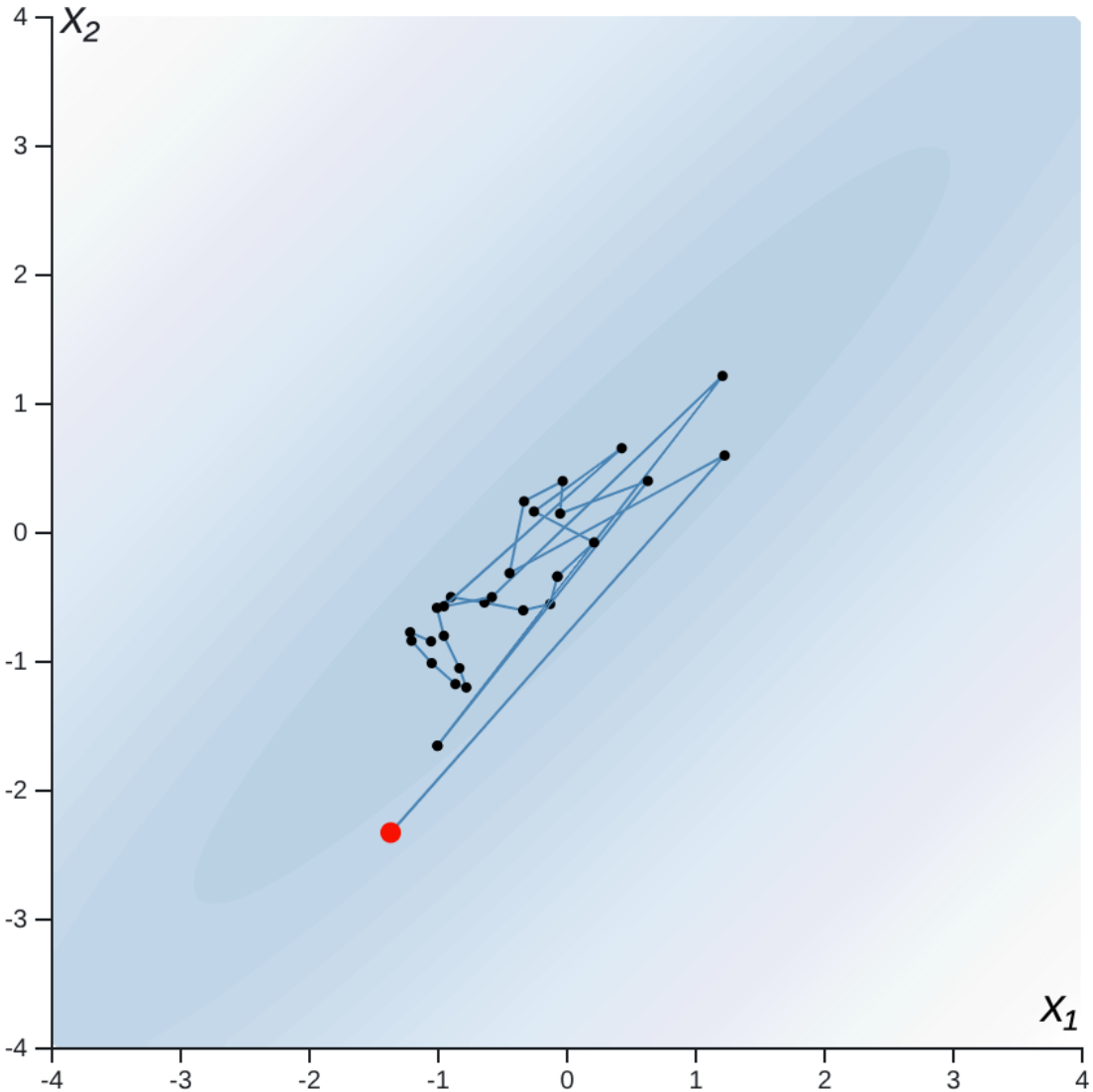
end

Output: draws $\boldsymbol{\theta}^{(b+1)}, \ldots, \boldsymbol{\theta}^{(b+m)}$ from $p(\boldsymbol{\theta}|\mathbf{y})$.

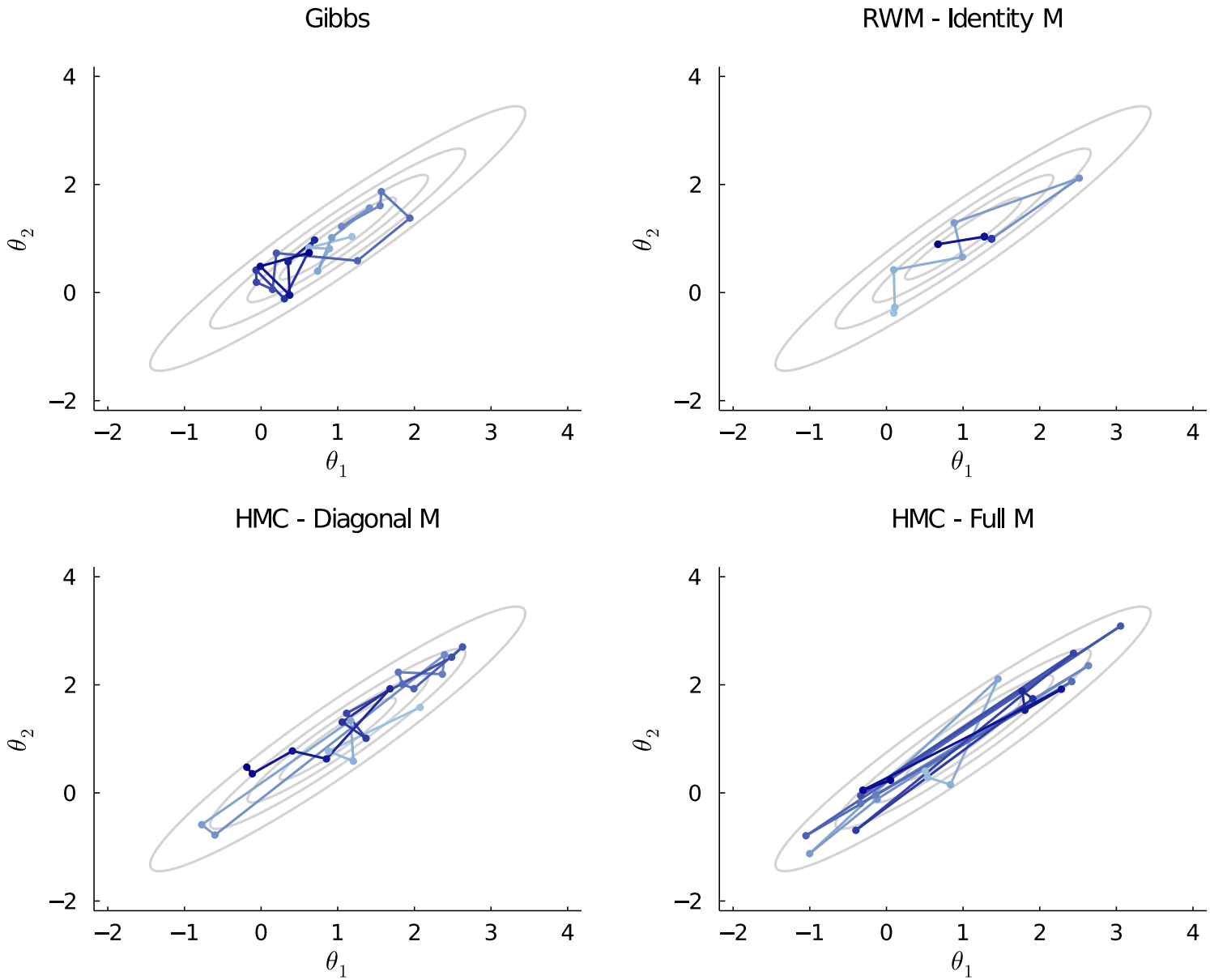
Tuning Hamiltonian Monte Carlo

- HMC is very efficient, but **needs careful tuning** to work.
- **Tuning parameters**:
  - **stepsize** ε,
  - **number of leapfrog** iterations L and
  - **mass matrix** M. (hello $J_{\mathbf{x}}^{-1}(\tilde{\boldsymbol{\theta}})$, our old friend)
- **No U-turn** sampler:
  - **Warm-up** to determine ε and L to get good acceptance rate.
  - Avoids U-turns in the Hamiltonian proposals.
- Drawbacks of HMC:
  - Need to **evaluate gradient of log posterior** many times during Hamiltonian iterations. Costly! (Subsampling HMC).
  - Difficulty with **multimodality** (true for most algorithms).
  - Standard HMC cannot handle **discrete parameters**. Mixture example. Some recent progress.

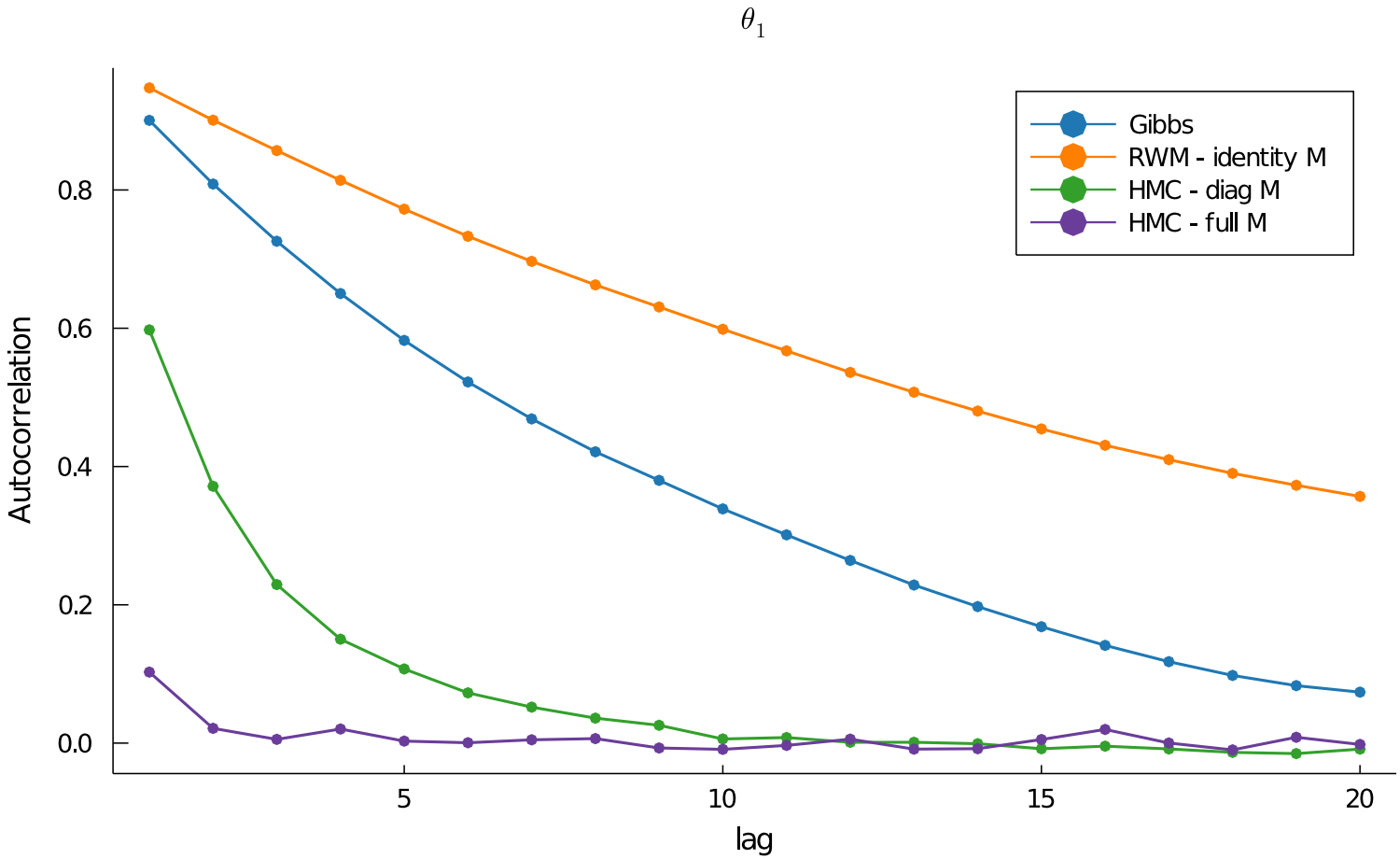
Interactive - HMC



Comparing algorithms for bivariate normal



Comparing algorithms for bivariate normal



Variational Inference

- Approximate posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$ with (simpler) distribution $q(\theta)$.
- Before: **Normal approximation** from optimization:

$$q(\boldsymbol{\theta}) = N\left[\tilde{\boldsymbol{\theta}}, J_{\mathbf{x}}^{-1}(\tilde{\boldsymbol{\theta}})\right]$$

- **Mean field Variational Inference (VI)**:

$$q(\boldsymbol{\theta}) = \prod_{i=1}^{p} q_i(\theta_i)$$

- **Parametric VI**: Parametric family $q_{\lambda}(\boldsymbol{\theta})$ with parameters λ. Example: $q(\boldsymbol{\theta}) = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. $\boldsymbol{\lambda} = (\boldsymbol{\mu}, \mathrm{Chol}(\Sigma))$.
- Find $q(\boldsymbol{\theta})$ that **minimizes the Kullback-Leibler divergence** between the true posterior p and the approximation q:

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

Mean field approximation

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

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

- **No specific functional forms** are assumed for the $q_j(\theta_j)$.
- **Optimal densities** can be shown to satisfy:

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

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

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

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

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

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

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

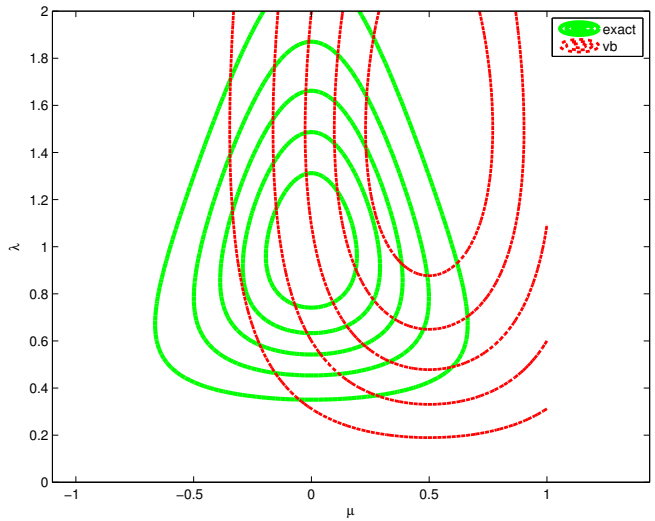
$$\tilde{\mu}_n = \tilde{w}\bar{x} + (1 - \tilde{w})\mu_0,$$

where

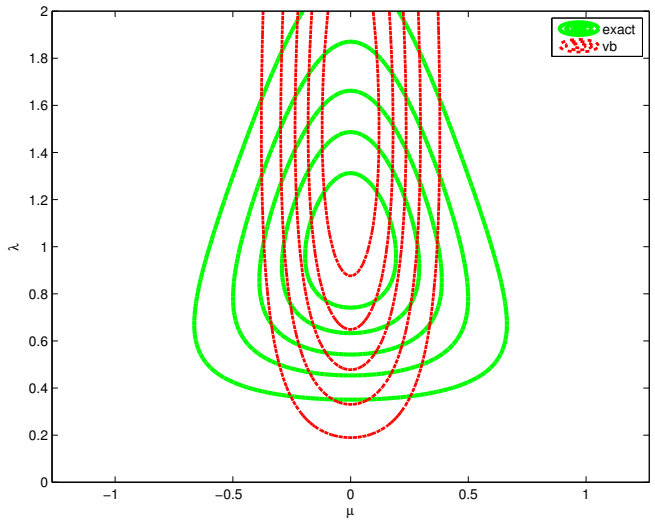
$$\tilde{w} = \frac{\frac{n}{\tilde{\sigma}_n^2}}{\frac{n}{\tilde{\sigma}_n^2} + \frac{1}{\tau_0^2}}$$

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

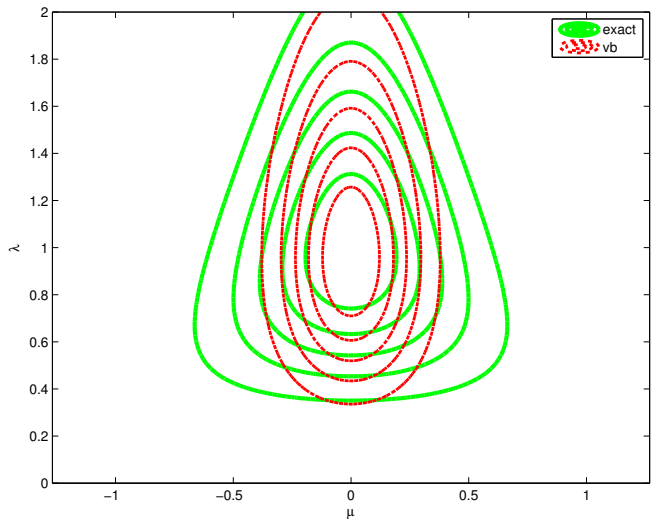
Initial values



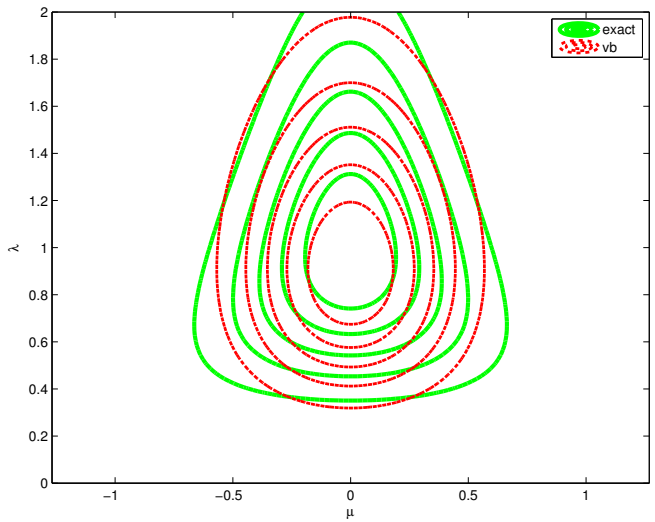
After updating q_μ



After updating q_{σ^2}



At convergence



Probit regression

- **Model**:

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

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

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

and

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

- Factorized **variational approximation**

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

VI for probit regression

VI posterior

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

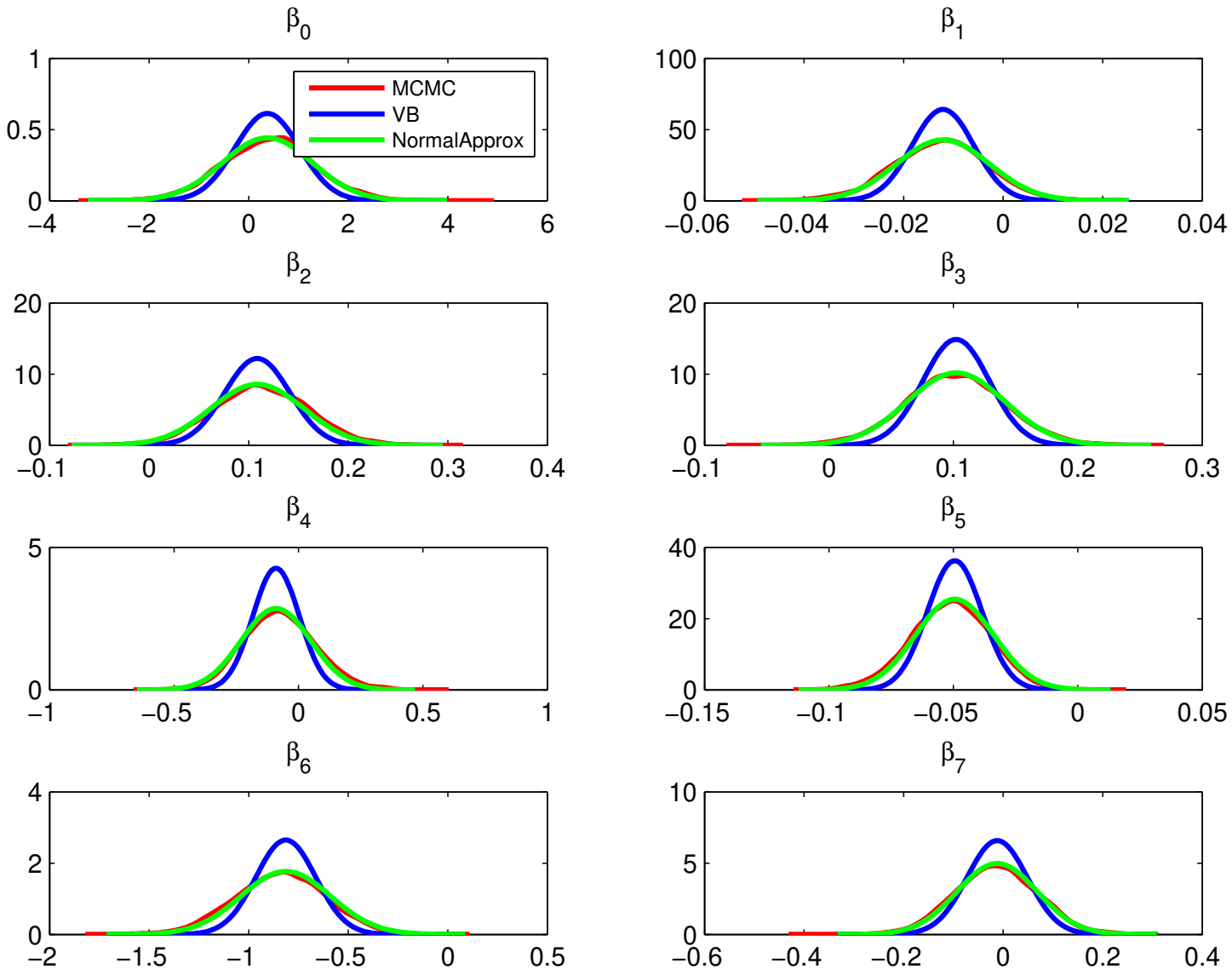
where

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

and

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

Probit example (n=200 observations)



Probit example

- Minimizing KL divergence = maximizing **evidence lower bound** (**ELBO**)

$$\text{ELBO} = E_q\left(\log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}\right)$$

- So proportional form of the posterior is enough.
- Can use ELBO to detect convergence.