

STATS 230: Computational Statistics

Advanced MCMC Methods

Babak Shahbaba

Department of Statistics, UCI

- Simple Metropolis algorithm and Gibbs sampler explore the posterior distribution using a random walk.
- While this strategy might work well for low-dimensional distributions, it could become very inefficient (e.g., high autocorrelation, missing isolated modes) for high-dimensional distributions.
- In this lecture, we discuss two useful strategies to improve the efficiency of Markov chain Monte Carlo methods.
- We are going to briefly discuss
 - ▶ Overrelaxation for Gibbs sampling
 - ▶ Hamiltonian Monte Carlo
- For more information on overrelaxation refer to Neal (1998). For Hamiltonian Monte Carlo, refer to Neal (2010).

- This method is used to improve Gibbs samplers by reducing the random walk behavior.
- Adler (1981) first introduced this method for situations where all the conditional distributions are Gaussian.
- Suppose we want to sample from the distribution of x_1, \dots, x_N , where

$$x_j | x_{-j} \sim N(\mu_j, \sigma_j^2)$$

Here, x_{-j} represents all random variables excluding x_j , and (μ_j, σ_j^2) are the parameters of the conditional distribution given all other parameters.

- Note that in Bayesian analysis the random variable of interest are model parameters and generally want to sample from their joint posterior distribution.

- Recall that Gibbs sampling sequentially updates $\{x_j\}$ by sampling from their conditional distribution $N(\mu_j, \sigma_j^2)$ given the current values of all other parameters

$$\begin{aligned}x_j^{(i+1)} &= \mu_j + \sigma_j z \\ z &\sim N(0, 1)\end{aligned}$$

- In Adler's method, we instead update x_j as follows:

$$\begin{aligned}x_j^{(i+1)} &= \mu_j + \alpha(x_j^{(i)} - \mu_j) + \sigma_j(1 - \alpha^2)^{1/2} z \\ z &\sim N(0, 1)\end{aligned}$$

where α is a constant usually slightly bigger than -1, for example -0.9. (In general, to make this method work we must have $-1 < \alpha < 1$.)

- Note that when $\alpha = 0$ this method reduces to the simple Gibbs sampling.

- Let's look at a simple example to see how overrelaxation works.
- Assume that we want to sample from the following distribution:

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.99 \\ 0.99 & 1 \end{pmatrix}\right)$$

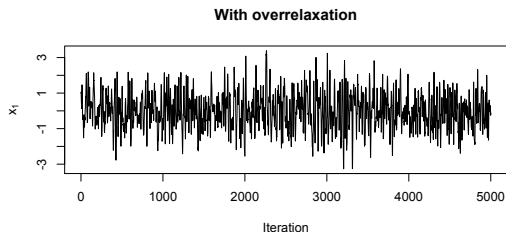
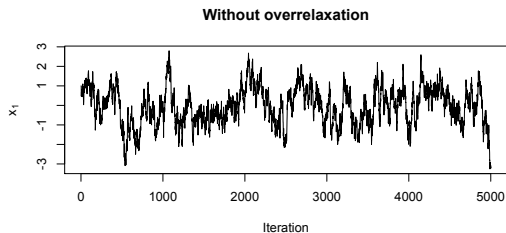
- Based on what we know about multivariate normal distributions, we have

$$x_1|x_2 \sim N(0.99x_2, (1 - 0.99^2))$$

$$x_2|x_1 \sim N(0.99x_1, (1 - 0.99^2))$$

Overrelaxation

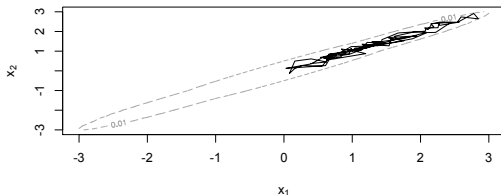
- The following graph shows the trace plots of samples for x_1 with and without overrelaxation



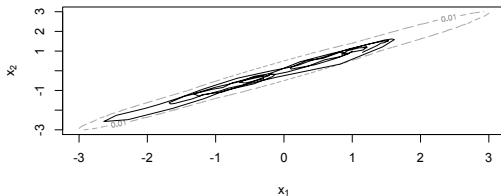
Overrelaxation

- The following graph shows the first 100 iterations, starting from (1, 1), with and without overrelaxation

Without overrelaxation



With overrelaxation



- Why overrelaxation works?
- It clearly leaves the distribution invariant: if

$$x_j^{(i)} | \cdot \sim N(\mu_j, \sigma_j^2)$$

then

$$x_j^{(i+1)} | \cdot \sim N(\mu_j, \sigma_j^2)$$

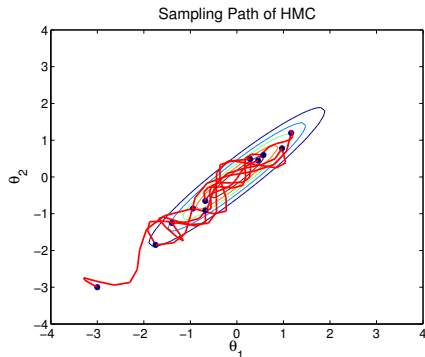
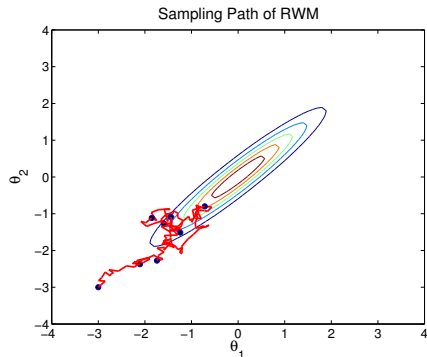
- Overrelaxation tends to propose states on the opposite side of the conditional mean compared to the current state.
- This causes the chain to move in a consistent direction so it explores the posterior distribution better.
- See Neal (1995) to more general forms of overrelaxation (<http://arxiv.org/pdf/bayes-an/9506004.pdf>)

Hamiltonian dynamics

- Hamiltonian dynamics is used to improve the efficiency of the Metropolis algorithm.
- It does this by reducing the random walk behavior through proposing distant states (from the current state) with high probability of acceptance (in theory, with probability 1).
- A little bit of history...
- In 1953, Metropolis et. al. introduced MCMC to simulate the distribution of states for a system of idealized molecule.
- In 1959, Alder and Wainwright introduced a system called *Hamiltonian dynamics* to simulate motion of molecules deterministically based on Newton's law of motion.
- In 1987, Duane et. al. combine the MCMC and the Hamiltonian dynamics. They called their method *Hybrid Monte Carlo* (HMC).
- The abbreviation HMC has also been used for *Hamiltonian Monte Carlo*.

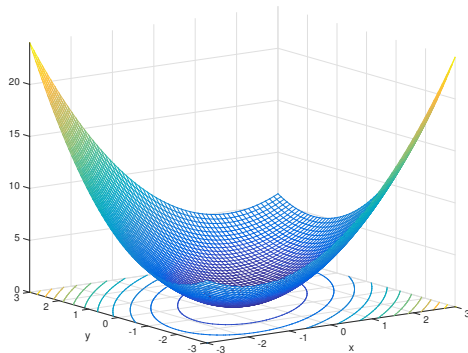
Hamiltonian Monte Carlo

- HMC proposes states that are distant from the current state, but nevertheless have a high probability of acceptance.



Hamiltonian Monte Carlo

- The sampler is viewed as a dynamic system moving on a surface defined by the *energy* function U : negative log density of the target distribution



- Distant proposals are found by numerically simulating Hamiltonian dynamics for some specified amount of fictitious time

Posterior sampling

- For Bayesian inference, posterior distribution is the target distribution

Potential energy

$$U(\theta) = - \sum_{i=1}^N \log P(y_i|\theta) - \log P(\theta)$$

- We augment the parameter space with fictitious momentum variables

Kinetic energy

$$K(p) = \frac{1}{2} p^\top M^{-1} p$$

- Define the Hamiltonian function $H(\theta, p) = U(\theta) + K(p)$
- The joint density of (θ, p) is

$$P(\theta, p) \propto \exp\{-H(\theta, p)\} = \exp\{-U(\theta)\} \cdot \exp\{-K(p)\}$$

- The marginal distribution of θ is the posterior distribution

Hamiltonian Dynamics

- We can generate a proposal by starting from the current state at time 0 and moving to the state at time t :

$$(\theta, p) = (\theta^{(0)}, p^{(0)}) \xrightarrow{\text{HD}} (\theta^{(t)}, p^{(t)}) = (\theta^*, p^*)$$

- *Hamilton's equations* determine how θ and p change over [fictitious] time

Hamilton's equations

$$\begin{aligned}\frac{d\theta_j}{dt} &= \frac{\partial H}{\partial p_j} = [M^{-1}p]_j \\ \frac{dp_j}{dt} &= -\frac{\partial H}{\partial \theta_j} = -\frac{\partial U}{\partial \theta_j}\end{aligned}$$

- Important properties (see <http://arxiv.org/pdf/1206.1901.pdf>):
 - ▶ **Reversibility**: the target distribution remain invariant
 - ▶ **Volume preservation**: the Jacobin determinant is 1
 - ▶ **Conservation of Hamiltonian**: the acceptance rate is one; θ^* is the next sample if HD is analytically solvable

Numerical Integration

- Numerical integration is employed when analytic solution is not available

Leapfrog

$$p_j(t + \varepsilon/2) = p_j(t) - (\varepsilon/2) \frac{\partial U}{\partial \theta_j}(\theta(t))$$

$$\theta_j(t + \varepsilon) = \theta_j(t) + \varepsilon \frac{\partial K}{\partial p_j}(p(t + \varepsilon/2))$$

$$p_j(t + \varepsilon) = p_j(t + \varepsilon/2) - (\varepsilon/2) \frac{\partial U}{\partial \theta_j}(\theta(t + \varepsilon))$$

- **Important properties:**

- ▶ **Stability:** numerically stable if ε is appropriately chosen
- ▶ **Reversibility and Volume preservation:** still hold
- ▶ **Conservation of Hamiltonian:** broken, but can be corrected by MH correction step with acceptance rate

$$\alpha = \min[1, \exp(-H(\theta^*, p^*) + H(\theta, p))]$$

Algorithm 1 HMC algorithm

Initialize $\theta^{(0)} = \text{current } \theta$

Sample new momentum $p^{(0)} \sim \mathcal{N}(0, M = I)$

Calculate current $H(\theta^{(0)}, p^{(0)}) = U(\theta^{(0)}) + \frac{1}{2}(p^{(0)})^\top p^{(0)}$

for $\ell = 1$ to L (leapfrog steps) **do**

$$p^{(\ell+\frac{1}{2})} = p^{(\ell)} - \varepsilon/2 \nabla_{\theta} U(\theta^{(\ell)})$$

$$\theta^{(\ell+1)} = \theta^{(\ell)} + \varepsilon p^{(\ell+\frac{1}{2})}$$

$$p^{(\ell+1)} = p^{(\ell+\frac{1}{2})} - \varepsilon/2 \nabla_{\theta} U(\theta^{(\ell+1)})$$

end for

Accept or reject according to the Metropolis acceptance probability

Example: logistic regression with $N(0, \sigma^2 I)$ prior

$$\nabla_{\beta_j} U(\beta) = - \sum_{i=1}^N \left[y_i - \frac{\exp(x_i \beta)}{1 + \exp(x_i \beta)} \right] x_{ij} + \beta_j / \sigma^2$$

A special case: Langevin Monte Carlo

- A special case: $L = 1$ and $M = I$
- This is called Langevin Monte Carlo,

Langevin dynamics

$$\begin{aligned}\theta^* &= \theta - \frac{\varepsilon^2}{2} \nabla_{\theta} U(\theta) + \varepsilon p \\ p^* &= p - \frac{\varepsilon}{2} \nabla_{\theta} U(\theta) - \frac{\varepsilon}{2} \nabla_{\theta} U(\theta^*)\end{aligned}$$

- Alternatively, we could ignore the momentum variable p and use the following asymmetrical proposal with MH acceptance probability

$$\theta^* \sim N\left(\theta - \frac{\varepsilon^2}{2} \nabla_{\theta} U(\theta), \varepsilon^2 I\right)$$

- Dropping the accept/reject step leads to an approximate Langevin method (see Neal, 1993)

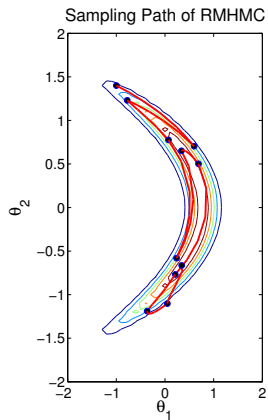
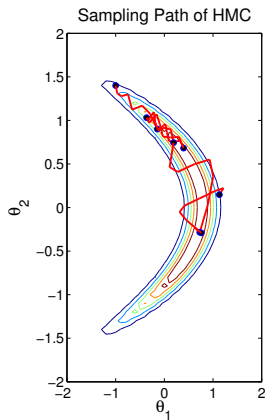
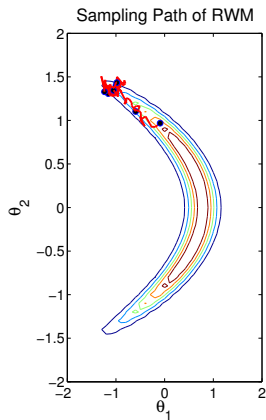
A more general case: Riemannian Manifold HMC

- Girolami and Calderhead (2011) have introduced a new method, called Riemannian Manifold HMC (RMHMC)
- They argue that it is more natural to put the Hamiltonian dynamic on Riemannian manifold of distributions rather than Euclidean space
- They follow Amari (2000) and use the Fisher information matrix, $G(\theta) = E[\nabla_{\theta}^2 U(\theta)]$, as a metric on the manifold
- That is, they use position specific mass matrix, $M = G(\theta)$
- Example: logistic regression

$$G_{jk}(\beta) = \sum_{i=1}^N x_{ij} x_{ik} \frac{\exp(x_i \beta)}{[1 + \exp(x_i \beta)]^2}, \quad j \neq k$$

- We can explore the parameter space more efficiently by exploiting its geometric properties
- The resulting dynamics is non-separable so instead of the standard leapfrog method we need to use the *generalized* leapfrog method

HMC vs. RMHMC



- In recent years, computational methods based on mini-batches of data have been quite successful
 - ▶ The underlying assumption: there is redundancy in big data
 - ▶ The overall information can be retrieved from a small subset
 - ▶ We can approximate functions at low computational cost
- Welling and Teh (2011) used this approach (stochastic gradient) for Langevin dynamics using mini-batches of size n from N observations

$$\theta^* = \theta + \frac{\varepsilon^2}{2} (\nabla_{\theta} P(\theta) + \frac{N}{n} \sum_{i=1}^n \nabla_{\theta} \log P(x_i|\theta)) + \varepsilon p$$

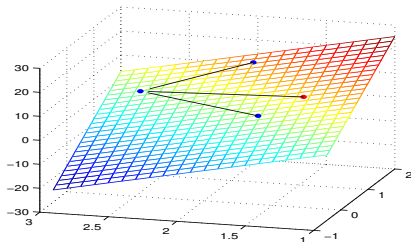
- They also dropped the accept/reject step

Precomputing strategies

- Finding optimum subsets by exploiting regularity in data space is difficult
- Using random subsets could lead to non-ignorable loss of information
- Therefore, (S. et al., 2011) previously proposed to identify a subset of influential points to split the Hamiltonian function (the likelihood part) into two parts
- Recently, Zhang et. al. (2015) have proposed to switch the focus of approximation from data space to parameter space (<http://arxiv.org/abs/1504.01418>)

A Toy Example – Gaussian Model

- Consider a Gaussian model with a Gaussian prior on the mean (no sampling required)
- The function value at any given point (red point) can be obtained by interpolation given a sample of three points (blue points)



$$\nabla_{\mu} U$$

Grid Approximation of HMC (GHMC)

$$\begin{aligned}\frac{dp_j}{dt} &= -\frac{\partial U}{\partial \theta_j} \\ \frac{d\theta_j}{dt} &= [M^{-1}p]_j\end{aligned}$$

Denote

Force

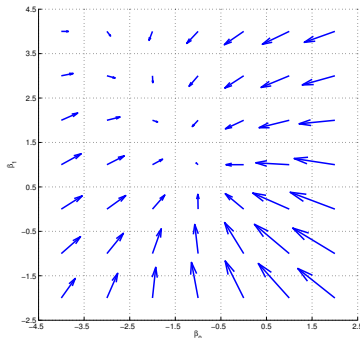
$$F = -\nabla U$$

- piecewise constant approximation

$$\tilde{F}(\theta) = F_{i,j} \triangleq F(c_{i,j}), \quad \text{if } \theta \in C_{i,j}$$

- piecewise linear approximation

$$\tilde{F}(\theta) = F_{i,j} + \nabla F_{i,j} \cdot (q - c_{i,j}), \quad \text{if } \theta \in C_{i,j}$$



Force map of a logistic regression model

True $\beta = (-1, 1)$