CS 5135/6035 Learning Probabilistic Models

Lecture 22: Markov Chain Monte Carlo Methods II

Gowtham Atluri

November 23, 2018

Reading Material

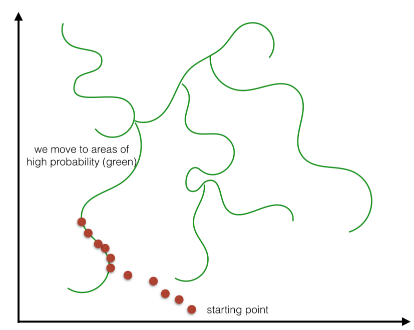
- Chapter 6. Metropolis-Hastings Algorithms
  Christian Robert and George Casella. Introducing Monte Carlo Methods with R
- Siddhartha Chib and Edward Greenberg. 'Understanding the Metropolis-Hastings algorithm.' The American Statistician, 1995.
- George Casella and Edward I. George. 'Explaining the Gibbs sampler.' The American Statistician 46, 1992.

Topics

- Markov Chain Monte Carlo Methods
- Metropolis-Hastings Algorithm
  - Random-walk Metropolis-Hastings
  - Independent Metropolis-Hastings
  - Choosing $q(x, y)$ or tuning
- Gibbs Sampling
  - Algorithm
  - Comparison with Metropolis-Hastings
  - As a special case of Metropolis-Hastings

Idea behind Markov Chain Monte Carlo Methods

- Instead of sampling i.i.d., sample from a Markov Chain



- *Markov Chain*- where we go next depends on our current state
- *Monte Carlo* - Simulating data

Metropolis-Hastings Algorithm (General Version)

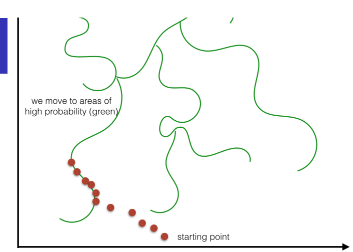
Algorithm:

1. Initialize $x_0 \sim q$
2. **for** iteration $i = 1, 2, \ldots$ **do**
3. Propose: $x_{cand} \sim q(x_i | x_{i-1})$
4. Acceptance Prob.:

$$\alpha(x_{cond} | x_{i-1}) = min\{1, \frac{q(x_{i-1}|x_{cand})f(x_{cand})}{q(x_{cand}|x_{i-1})f(x_{i-1})}\}$$

5. $u \sim Uniform(0, 1)$
6. **if** $u < \alpha$ **then**
7. Accept the proposal $x_i \leftarrow x_{cand}$
8. **else**
9. Reject the proposal $x_i \leftarrow x_{i-1}$
10. **end if**
11. **end for**

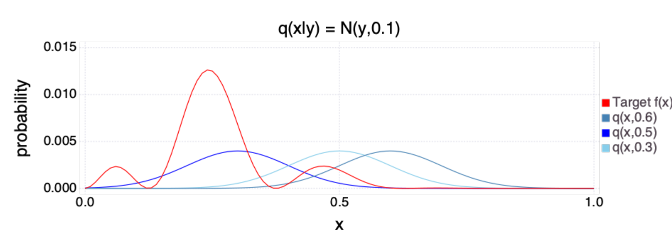
Example: MH Setup

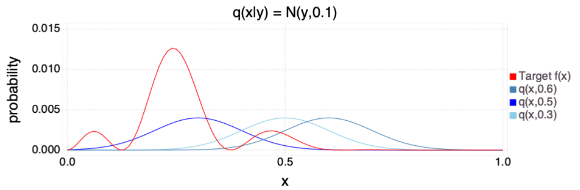


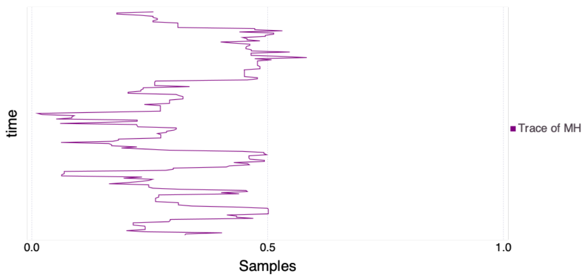
$$p(\theta|y) = f(x) = 2\theta^2(1-\theta)^8 cos^2(4\pi\theta)$$

Candidate/proposal distribution:

$$q(x_{cand}|x) = \mathcal{N}(x, 0.1)$$







Random Walk Metropolis-Hastings

- From our example, proposal distr. $q(x_{cand}|x) = \mathcal{N}(x, 0.1)$; $x_{cand} \sim \mathcal{N}(x, 0.1)$
  - Alternatively $x_{cand} = x + \epsilon$; $\epsilon \sim \mathcal{N}(0, 0.1)$
- More generally, $x_{cand} = x_{i-1} + \epsilon$
  - ϵ is a *random perturbation* with a distribution independent of current state
  - E.g., $x_{cand} = x_{i=1} + \epsilon_t$, where $\epsilon_t \sim Uniform(-\delta, \delta)$
  - E.g., $x_{cand} = x_{i=1} + \epsilon_t$, where $\epsilon_t \sim Normal(0, \tau^2)$
- In the context of the general Metropolis-Hastings algorithm
  - $q(x|y) = q(y - x)$
- Markov chain associated with q is a *radom walk*, when it is symmetric around 0, i.e, $q(-t) = q(t)$
  - due to acceptance step in M-H,M-H samples are *not* a random walk

Random Walk Metropolis-Hastings

- Acceptance probability

$$\alpha(x_{cond}|x_{i-1}) = min\{1, \frac{q(x_{i-1}|x_{cand})f(x_{cand})}{q(x_{cand}|x_{i-1})f(x_{i-1})}\} = min\{1, \frac{f(x_{cand})}{f(x_{i-1})}\}$$

- 'Uphill' proposals are always accepted
  - when $f(x_{cand}) > f(x_{i-1})$, $\alpha = 1$
- 'Downhill' proposals are accepted with probability equal to the relative 'heights' of the target at the proposed and current values.
  - When $f(x_{cand}) < f(x_{i-1})$, $\alpha = \frac{f(x_{cand})}{f(x_{i-1})}$
- The above simplification of α is not unique to random-walk M-H
  - If $q(x_{i-1}|x_{cand}) = q(x_{cand}|x_{i-1})$, $\alpha = min\{1, \frac{f(x_{cand})}{f(x_{i-1})}\}$

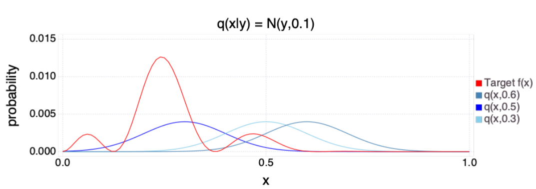
Choosing $q(x, y)$

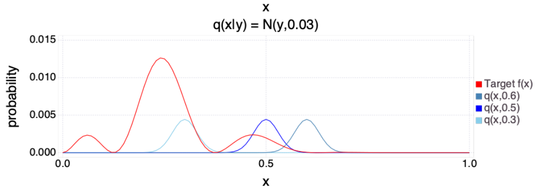
- The induced Markov chain should be irreducible, with short mixing time, to allow full coverage of the state-space
  - Support of q should include support of f ($support(f) \subset support(q)$)
- Typically $q(x|y)$ is selected from a family of distributions
  - that requires specification of location and scale parameters
  - E.g., Normal, Uniform, Cauchy, Laplace, Student's T-distribution
- A $q(x|y)$ with a small 'scale' will limit the step size of the Markov Chain

Example $p(\theta|y) = f(x) = 2\theta^2(1-\theta)^8 cos^2(4\pi\theta)$

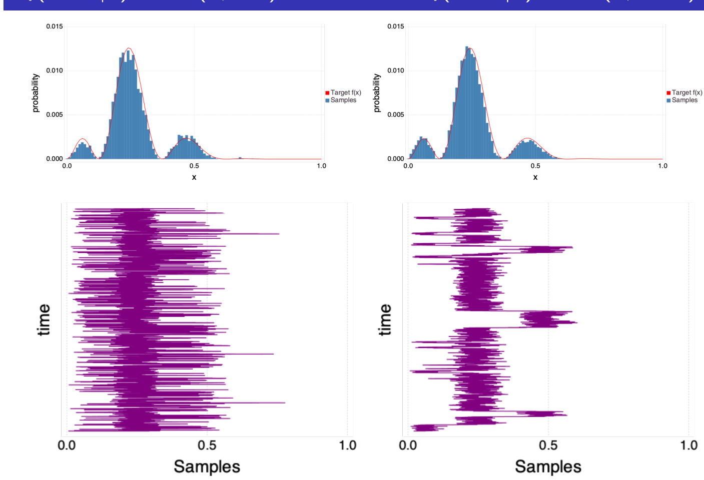
Candidate/proposal distribution:

$$q(x_{cand}|x) = \mathcal{N}(x, 0.1) \quad vs. \quad q(x_{cand}|x) = \mathcal{N}(x, 0.03)$$





$q(x_{cand}|x) = \mathcal{N}(x, 0.1)$ vs. $q(x_{cand}|x) = \mathcal{N}(x, 0.03)$



Choosing $q(x, y)$

- The induced Markov chain should be irreducible, with short mixing time, to allow full coverage of the state-space
  - Support of q should include support of f ($support(f) \subset support(q)$)
- Typically $q(x|y)$ is selected from a family of distributions
  - that requires specification of location and scale parameters
  - E.g., Normal, Uniform, Cauchy, Laplace, Student's T-distribution
- A $q(x|y)$ with a small 'scale' will limit the step size of the Markov Chain
- When we choose a $q(x|y)$ that is independent of the current state y
  - $q(x|y) = q(x)$
  - This is a special case of the Metropolis-Hastings Algorithm
    - Referred to as *independent* Metropolis-Hastings algorithm
    - appears to be a straightforward generalization of Accept-reject method

Independent Metropolis-Hastings $q(x|y) = q(x)$

- Choosing $q(x|y)$ that is independent of the current state y

Algorithm:

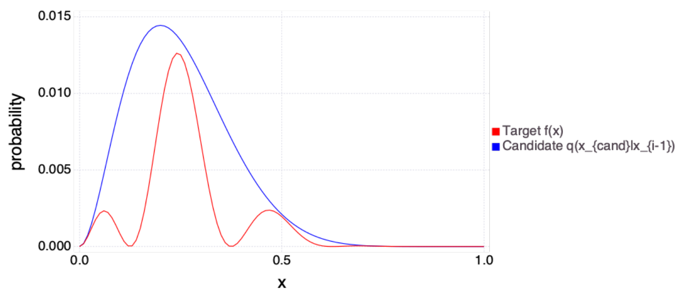
1. Initialize $x_0 \sim q$
2. **for** iteration $i = 1, 2, \ldots$ **do**
3. Propose: $x_{cand} \sim q(x_i)$
4. Acceptance Prob.:

$$\alpha(x_{cond}|x_{i-1}) = min\{1, \frac{q(x_{i-1})f(x_{cand})}{q(x_{cand})f(x_{i-1})}\}$$

5. $u \sim Uniform(0, 1)$
6. **if** $u < \alpha$ **then**
7. Accept the proposal $x_i \leftarrow x_{cand}$
8. **else**
9. Reject the proposal $x_i \leftarrow x_{i-1}$
10. **end if**
11. **end for**

Example: $q(x|y)$ is independent of current state y

Target distribution: $f(x) = 2\theta^2(1-\theta)^8 cos^2(4\pi\theta)$ Candidate/proposal distribution: $q(x) = Beta(3, 9)$



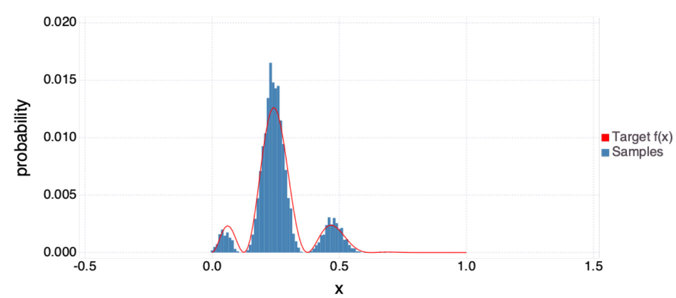
Independent MH vs. Accept-Reject Method

- Independent Metropolis-Hastings
  - appears to be a straightforward generalization of Accept-reject method
- Repeated occurrences
  - no repeated occurrences in Accept-Reject Method
  - repeated occurrences possible in Independent Metropolis-Hastings
    - Step **9**: Reject the proposal $x_i \leftarrow x_{i-1}$
- Samples are
  - i.i.d in Accept-Reject Method
  - Not i.i.d in Independent Metropolis-Hastings
- Determining upper bound M using $f(x)/g(x) \leq M$
  - required in Accept-Reject Method
  - not required in Independent Metropolis-Hastings

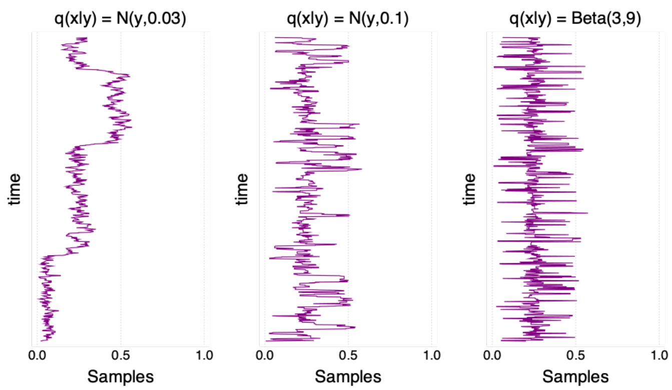
Independent Metropolis-Hastings

```
f(x) = 2.*x.^2.*(1.-x).^8.*cos.(4.*pi.*x).^2;
q(x) = pdf.(Beta(3,9),x);

function metropolis_hastings(n)
    x = zeros(n); count = 1;
    x[1] = rand(Beta(3,9));

    while(count < n)
        x_cand = rand(Beta(3,9));
        rho = (q(x[count])/q(x_cand))*(f(x_cand)/f(x[count]));
        alpha = minimum([1,rho]);
        u = rand();
        if (u < alpha)
            x[count] = x_cand;
        else
            count = count + 1
            x[count] = x[count-1];
        end
    end
    return x;
end
```


Independent Metropolis-Hastings



Independent Metropolis-Hastings



- When $q(x|y) = q(x)$, any state is possible from the current state

Choosing proposal density: observations

- The spread of the of the proposal density affects
  1. acceptance rate
  2. region of the sample space covered by the chain
- When the chain converged and density is sampled around the mode
  - If spread is extremely large, next sample will be far from current value
    - low probability of being accepted
  - If spread is too small, it will take too long to traverse support of target density
    - low probability regions will be undersampled
- Proposal density needs to be tuned appropriately

Integration using MCMC

- While the examples we considered involve 'sampling'
  - MCMC methods are suited for integration as well
- **Ergodic Theorem**: For a finite irreducible chain with stationary distribution π

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} h(x_t) = \mathbb{E}_\pi(h(x))$$

  - This expectation is the same as the integral $\int h(x)\pi(x)dx$
- **Approach:**
  - Draw n samples from $\pi(x)$ using Metropolis-Hastings
  - Compute the values for $h(x)$ using these samples
  - Compute the average of the $h(x)$ values

Gibbs sampling: Introduction

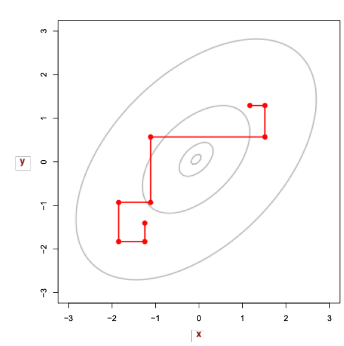
- Metropolis-Hastings is an MCMC method that generates samples by traversing the support of a target-distribution
  - If target distribution is high-dimensional
    - proposal distribution is also high-dimensional
    - selecting a suitable proposal that is not too narrow or too broad can be challenging
- We need a method of sample generation that did not demand artful tuning of a proposal distribution
- Gibbs sampling is one such method
  - obviates the need for a separate proposal distribution
  - makes other demands (to sample from conditional densities)
- Gibbs sampling is also an MCMC method
  - Geman and Geman 1984 paper on *Image processing models*
  - Sampling on a *Gibbs random field*, the name stuck

Gibbs sampling: Introduction

- Gibbs sampling allows us to generate samples from joint target density functions
  - Useful for sampling from a joint posterior $p(\theta_1, \theta_2, \ldots, \theta_d|y)$
- Gibbs sampling simplifies a complex high-dimensional problem
  - by breaking it down into simple, low-dimensional problems
- To draw samples from $f(x, y)$, Gibbs sampler draws from $f(x|y)$ and $f(y|x)$
  - Draw $x_{t+1} \sim f(x|y_t)$
  - Draw $y_{t+1} \sim f(y|x_t)$
  - Samples $x_0, y_0, x_1, y_1, \ldots, x_n, y_n$
- Assumes we can generate samples from $f(x|y)$ and $f(y|x)$

Gibbs sampling: Introduction

- To draw samples from $f(x, y)$
  - Draw $x_{t+1} \sim f(x|y_t)$
  - Draw $y_{t+1} \sim f(y|x_t)$
- Each step is parallel to one of the parameter axis
  - as only one component value is changed



Gibbs sampling: A general approach

Algorithm:

1. Initialize $x^{(0)} \sim q(x)$
2. **for** iteration $i = 1, 2, \ldots$ **do**
3. $\quad x_1^{(i)} \sim p(x_1 | x_2 = x_2^{(i-1)}, x_3 = x_3^{(i-1)}, \ldots, x_d = x_d^{(i-1)})$
4. $\quad x_2^{(i)} \sim p(x_2 | x_1 = x_1^{(i-1)}, x_3 = x_3^{(i-1)}, \ldots, x_d = x_d^{(i-1)})$

$$\vdots$$

5. $\quad x_d^{(i)} \sim p(x_d | x_2 = x_2^{(i-1)}, x_3 = x_3^{(i-1)}, \ldots, x_{d-1} = x_{d-1}^{(i-1)})$
6. **end for**

- GS assumes that we can draw samples from the full conditionals

Gibbs sampling vs. Metropolis-Hastings

- Both Gibbs and MH are MCMC methods
  - Both generates samples from a Markov Chain
- Sample generation
  - In MH, new candidate $\boldsymbol{x}^{cand} = \{x_1^{cand}, x_2^{cand}, \ldots, x_d^{cand}\}$ sampled from proposal distr.
  - At each step in Gibbs Sampling, one of the components is sampled
    - E.g., x_1^{cand} in step 1. x_2^{cand} in step 2, etc.
- Sampling distribution
  - In MH, is a proposal distribution that is selected by the user
  - Gibbs uses full conditional distribution
    - E.g., $x_1^{(i)} \sim p(x_1 | x_2 = x_2^{(i-1)}, x_3 = x_3^{(i-1)}, \ldots, x_d = x_d^{(i-1)})$
- Acceptance Probability
  - Computed as $\alpha = min\{1, \rho\}$ in MH
  - All samples are accepted in Gibbs ($\alpha = 1$)

Gibbs Sampling - a special case of MH

- Let x_i be the i^{th} variable and x_{-i} be all variables except x_i
- Let $p(x_1, \ldots, x_d)$ be the target distribution we want to simulate
- Let $Q(x_i', x_{-i} | x_i, x_{-i}) = \frac{1}{k} p(x_i' | x_{-i})$
  - because at each step, we are drawing $x_i' \sim p(x_i' | x_{-i})$
- Let $\alpha(x_i', x_{-i} | x_i, x_{-i}) = min(1, \rho)$, where

$$\begin{aligned}
\rho &= \frac{q(x_{i-1} | x_{cand}) f(x_{cand})}{q(x_{cand} | x_{i-1}) f(x_{i-1})} &&= \frac{Q(x_i, x_{-i} | x_i', x_{-i})}{Q(x_i', x_{-i} | x_i, x_{-i})} \frac{p(x_i', x_{-i})}{p(x_i, x_{-i})} \\
&= \frac{p(x_i', x_{-i})}{p(x_i, x_{-i})} \frac{p(x_i | x_{-i})}{p(x_i' | x_{-i})} &&= \frac{p(x_i' | x_{-i}) p(x_{-i})}{p(x_i | x_{-i}) p(x_{-i})} \frac{p(x_i | x_{-i})}{p(x_i' | x_{-i})} \\
&= 1
\end{aligned}$$

- Hence, acceptance probability $\alpha = 1$

Example

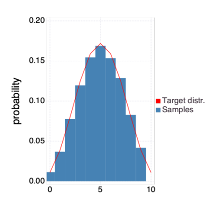
- Let $x|\theta \sim Binomial(n, \theta)$, $\theta \sim Beta(a, b)$, where $n = 10, a = 5, b = 5$
- Joint distribution

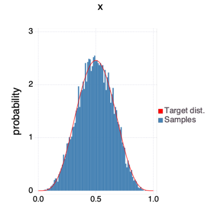
$$f(x, \theta) = \binom{n}{x} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{x+a-1} (1-\theta)^{n-x+b-1}$$

- We want to generate samples from this joint distr. $f(x, \theta)$
- To use Gibbs sampling we need conditionals $f(x|\theta)$ and $f(\theta|x)$
- Using, $f(\theta|x) = f(x, \theta)/f(x)$, we have $f(\theta|x) = Beta(x + a, n - x + b)$
- Approach
  - Initialize θ^0, x^0
  - Iterate from $i = 1$ to N
    - $\theta^i \sim f(\theta|x)$
    - $x^i \sim f(x|\theta)$

Julia code

```
N = 10000; n = 10; a = 5; b =5;
theta = zeros(N);
x = zeros(N);
theta[1] = rand(Beta(a,b));
x[1] = rand(Binomial(n,theta[1]));
for i=2:N
  theta[i] = rand(Beta(x[i-1]+a,n-x[i-1]+b));
  x[i] = rand(Binomial(n,theta[i-1]));
end
```





- As we sampled (x, θ) from the joint, we can use corresponding components from these sequences to approximate the marginal distributions
  - $x \sim BetaBinomial(n, a, b)$
  - $\theta \sim Beta(a, b)$

Example

- Let $x|\theta \sim Binomial(n, \theta)$, $\theta \sim Beta(a, b)$, where $n = 10, a = 5, b = 5$
- Joint distribution

$$f(x, \theta) = \binom{n}{x} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{x+a-1} (1-\theta)^{n-x+b-1}$$

- Here from $f(x|\theta)$ and $f(\theta)$, we were able to write the joint distribution
- Gibbs sampling is indispensable when $f(x, y)$, $f(x)$, and $f(y)$ cannot be calculated

$$f(x|y) \propto y e^{-yx}, \quad 0 < x < k < \infty$$
$$f(y|x) \propto x e^{-xy}, \quad 0 < y < k < \infty$$

  - Where k is some known constant
  - Restriction to the interval (0, B) ensures that marginal $f(x)$ exist
  - The form of this marginal is not easily calculable
  - Gibbs sampling can approximate $f(x, y)$, $f(x)$, and $f(y)$

Summary

- Metropolis-Hastings approaches
  - Random-walk MH
  - Independent MH
- Choosing the right proposal distribution is key to MH
  - acceptance rate will be low for a ‘too narrow’ proposal distr.
  - coverage of the target support will be slow for a ‘too broad’ proposal distr.
- Gibbs sampling
  - Overcomes the limitation of MH
  - No need to choose a proposal distribution
  - Assumes full conditionals can be simulated
  - Approximates joint and marginals
    - Even when they cannot be calculated