LECTURE 20: OVERVIEW OF MARKOV CHAIN MONTE CARLO

STAT 598z: Introduction to computing for statistics

Vinayak Rao

Department of Statistics, Purdue University

April 8, 2019

Markov chain Monte Carlo

Recall Monte Carlo: produce independent samples from $p(x)$, and use sample averages to approximate expectations.

In high dims, hard to even sample from $p(x)$!

Markov chain Monte Carlo

Recall Monte Carlo: produce independent samples from $p(x)$, and use sample averages to approximate expectations.

In high dims, hard to even sample from $p(x)$!

Rather than making independent proposals, exploit previous proposals to make good proposals

Allows us to find and explore useful regions of X-space

Markov chain Monte Carlo

Recall Monte Carlo: produce independent samples from $p(x)$, and use sample averages to approximate expectations.

In high dims, hard to even sample from $p(x)$!

Rather than making independent proposals, exploit previous proposals to make good proposals

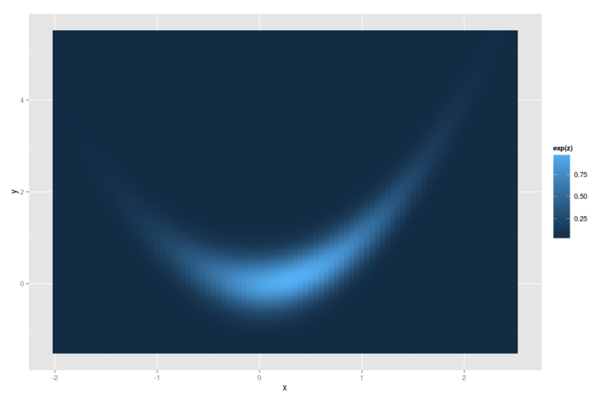
Allows us to find and explore useful regions of X-space

Simplest case: use current proposal to make a new proposal

The resulting algorithm: Markov chain Monte Carlo.

(A Markov chain: future independent of past given present)

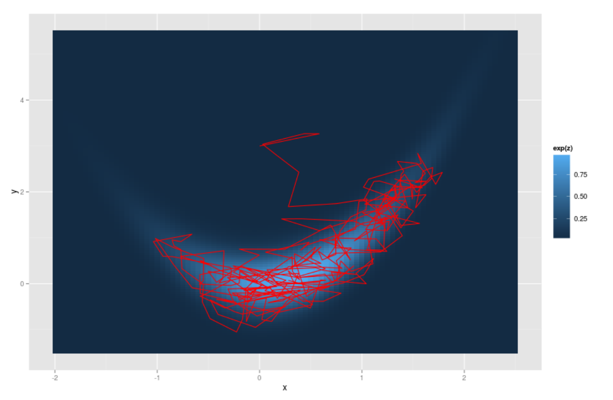
Markov chain Monte Carlo



The Rosenbrock density (a.k.a. the banana density)

$$p(x, y) \propto \exp\left(-(a - x)^2 - b(y - x^2)^2\right) \quad \text{(here } a = .3, b = 3\text{)}$$

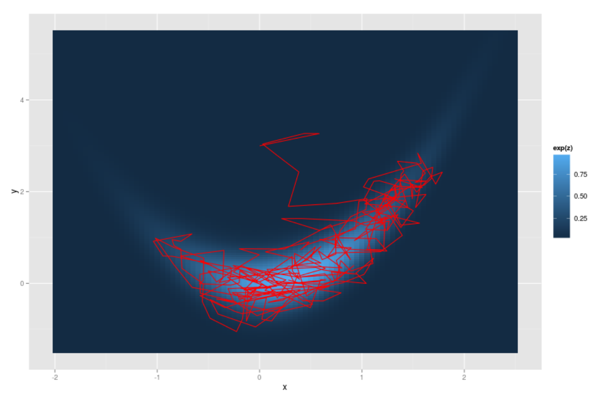
Markov chain Monte Carlo



A random walk:

- start somewhere arbitrary
- make local moves

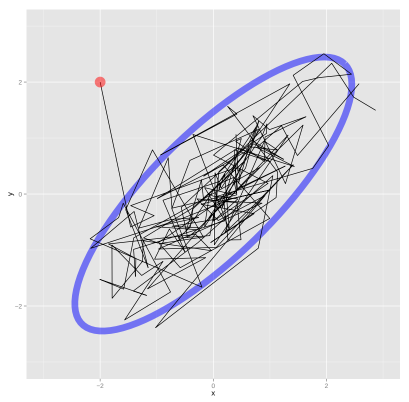
Markov chain Monte Carlo



- Discard initial ‘burn-in’ samples
- Use remaining to obtain Monte Carlo estimates:

$$\frac{1}{N}\sum_{i=1}^{N} f(x_i) \approx \mathbb{E}_p[f]$$

Markov chain Monte Carlo



A random walk over a 2-d Gaussian

- The goal of MCMC if to find a set of local moves that produce samples (asymtotically) from the right distribution

Markov chain Monte Carlo

- The goal of MCMC if to find a set of local moves that produce samples (asymtotically) from the right distribution
- The art of MCMC is to find local moves than coverge rapidly (a chain that 'mixes rapidly')

MCMC

Let element x_i of the chain have distribution p_i

Write $T(x_i \to x_{i+1}) = p(x_{i+1}|x_i)$ for the transition kernel of the chain.

Then, the $(i+1)$st element has distribution

$$p_{i+1}(x_{i+1}) = \int T(x_i \to x_{i+1}) p_i(x_i) \mathrm{d}x_i$$

p is the stationary/equilibrium distribution of the Markov chain if

$$p(x') = \int T(x \to x') p(x) \mathrm{d}x$$

MCMC: a first look

For a transition function $T(\cdot \to \cdot)$ with stationary distribution p

- Initialize x_0 from some distribution p_0
- Run a Markov chain for $(B+N)$ iterations with transition T

All x_i for $i > B$ are approximately distributed as p

MCMC: A FIRST LOOK

For a transition function $T(\cdot \to \cdot)$ with stationary distribution p

- Initialize x_0 from some distribution p_0
- Run a Markov chain for $(B + N)$ iterations with transition T

All x_i for $i > B$ are approximately distributed as p

- Discard the first B 'burn-in' samples
- Calculate Monte Carlo average with remaining N samples

$$\frac{1}{N} \sum_{i=B+1}^{B+N} f(x_i) \approx \mathbb{E}_p[f]$$

MCMC: A FIRST LOOK

For a transition function $T(\cdot \to \cdot)$ with stationary distribution p

- Initialize x_0 from some distribution p_0
- Run a Markov chain for $(B + N)$ iterations with transition T

All x_i for $i > B$ are approximately distributed as p

- Discard the first B 'burn-in' samples
- Calculate Monte Carlo average with remaining N samples

$$\frac{1}{N} \sum_{i=B+1}^{B+N} f(x_i) \approx \mathbb{E}_p[f]$$

Markov chain Monte Carlo estimate

MCMC

We want to sample from a probability distribution $p(x) = \frac{f(x)}{Z}$

How do we design an appropriate transition kernel $T(\cdot \to \cdot)$?

MCMC

We want to sample from a probability distribution $p(x) = \frac{f(x)}{Z}$

How do we design an appropriate transition kernel $T(\cdot \to \cdot)$?

Different MCMC algorithms take different approaches

The simplest is the Metropolis-Hastings algorithm

Metropolis-Hastings

Metropolis-Hastings (MH):

- Let current state be x_i
- Propose a new state from $q(w|x_i)$

If we set $x_{i+1} = w$ the resulting Markov chain will have the wrong stationary distribution

Metropolis-Hastings (MH):

- Let current state be x_i
- Propose a new state from $q(w|x_i)$

If we set $x_{i+1} = w$ the resulting Markov chain will have the wrong stationary distribution

- Instead, set $x_{i+i} = w$ (accept) with probability

$$\min(1, \frac{p(w)q(x_i|w)}{p(x_i)q(w|x_i)}) = \min(1, \frac{f(w)q(x_i|w)}{f(x_i)q(w|x_i)})$$

  Otherwise, set $x_{i+i} = x_i$ (reject)

Metropolis-Hastings

Metropolis-Hastings (MH):

- Let current state be x_i
- Propose a new state from $q(w|x_i)$

If we set $x_{i+1} = w$ the resulting Markov chain will have the wrong stationary distribution

- Instead, set $x_{i+i} = w$ (accept) with probability

$$\min(1, \frac{p(w)q(x_i|w)}{p(x_i)q(w|x_i)}) = \min(1, \frac{f(w)q(x_i|w)}{f(x_i)q(w|x_i)})$$

  Otherwise, set $x_{i+i} = x_i$ (reject)

Under mild conditions, this corrected Markov chain has the right stationary distribution

Works for any choice of q so long as it's possible to get from any part of space to any other (eventually)

Metropolis-Hastings

Works for any choice of q so long as it's possible to get from any part of space to any other (eventually)

Acceptance probability:

$$\min(1, \frac{p(y)q(x_i|y)}{p(x_i)q(y|x_i)}) = \min(1, \frac{f(y)q(x_i|y)}{f(x_i)q(y|x_i)})$$

We just have to evaluate the target density $p(x) = \frac{f(x)}{Z}$ up to a proportionality constant

Don't need the normalization constant Z!

Metropolis-Hastings

Works for any choice of q so long as it's possible to get from any part of space to any other (eventually)

Acceptance probability:

$$\min(1, \frac{p(y)q(x_i|y)}{p(x_i)q(y|x_i)}) = \min(1, \frac{f(y)q(x_i|y)}{f(x_i)q(y|x_i)})$$

We just have to evaluate the target density $p(x) = \frac{f(x)}{Z}$ up to a proportionality constant

Don't need the normalization constant Z!

We only need to

- sample from q
- evaluate transition probabilities $q(y|x)$
- evaluate the target density upto a normalization constant

Choice of proposal distribution q

- Common choice is a Gaussian centered at previous sample:

$$w | x_i \sim \mathcal{N}(x_i, \sigma^2)$$

- Equivalently,

$$w = x_i + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Choice of proposal distribution q

- Common choice is a Gaussian centered at previous sample:

$$w|x_i \sim \mathcal{N}(x_i, \sigma^2)$$

- Equivalently,

$$w = x_i + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

For this proposal $q(w|x_i) = q(x_i|w)$

$$\min(1, \frac{f(w)q(x_i|w)}{f(x_i)q(w|x_i)}) = \min\left(1, \frac{f(w)}{f(x_i)}\right)$$

Choice of proposal distribution q

- Common choice is a Gaussian centered at previous sample:

$$w|x_i \sim \mathcal{N}(x_i, \sigma^2)$$

- Equivalently,

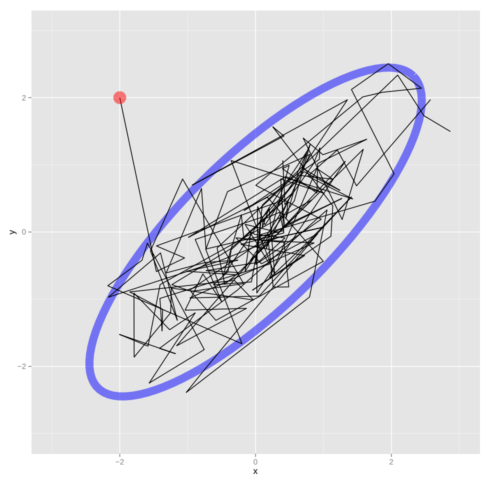
$$w = x_i + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

For this proposal $q(w|x_i) = q(x_i|w)$

$$\min(1, \frac{f(w)q(x_i|w)}{f(x_i)q(w|x_i)}) = \min\left(1, \frac{f(w)}{f(x_i)}\right)$$

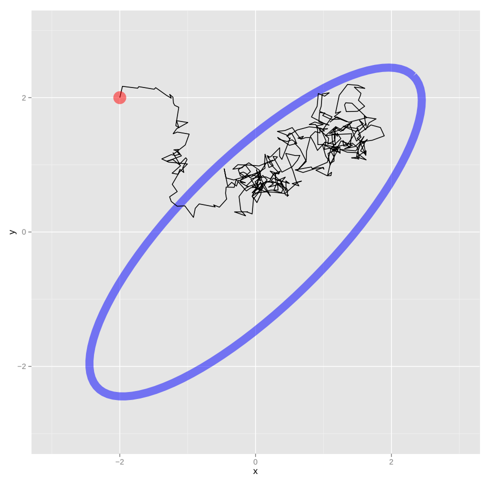
- always accept better proposals
- sometimes accept worse proposals

How do we chose the proposal variance?



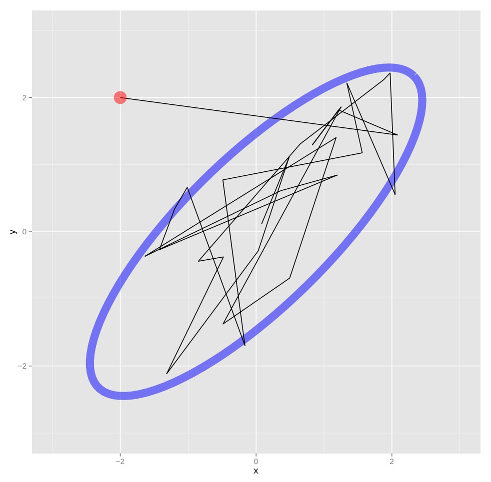
$$\sigma^2 = 1$$

How do we chose the proposal variance?



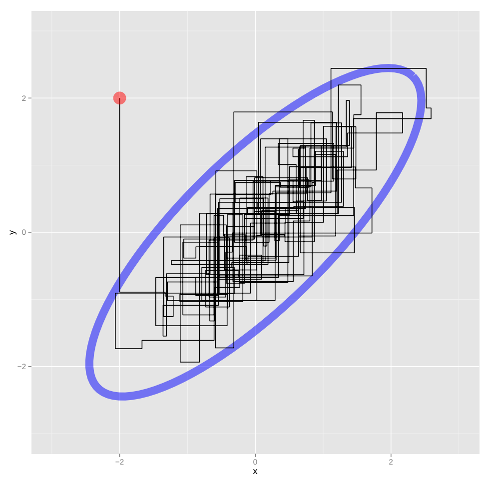
$$\sigma^2 = .1$$

How do we chose the proposal variance?



$$\sigma^2 = 5$$

Sample one component at a time



Gibbs sampling

Consider a set of variables $(x(1), \cdots, x(d))$

Gibbs sampling cycles though these sequentially (or randomly)

At the *i*th step, it updates $x(i)$ conditioned on the the rest:

$$x(i) \sim p(x(i)|x(1), \ldots, x(i-1), x(i+1), \ldots, x(n))$$

Often these 1-d conditionals are much simpler than the joint

Think of coordinate descent