LECTURE 15: MARKOV CHAIN MONTE CARLO

STAT 545: Introduction to computational statistics

Vinayak Rao

Department of Statistics, Purdue University

November 1, 2016

Markov chain Monte Carlo

For rejection/importance sampling proposal distribution must be similar to the distribution of interest

In high dims, hard to find reasonable proposal distributions

Markov chain Monte Carlo

For rejection/importance sampling proposal distribution must be similar to the distribution of interest

In high dims, hard to find reasonable proposal distributions

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

For rejection/importance sampling proposal distribution must be similar to the distribution of interest

In high dims, hard to find reasonable proposal distributions

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

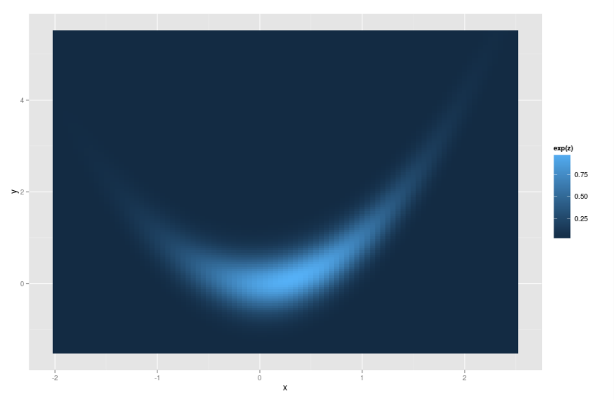
Allows us to find and explore useful regions of X-space

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)

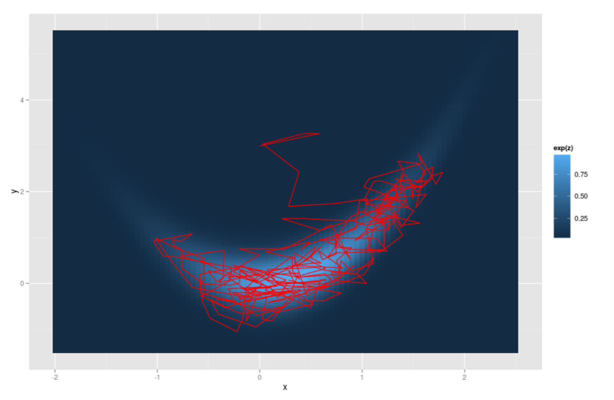
Markov chain Monte Carlo



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

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

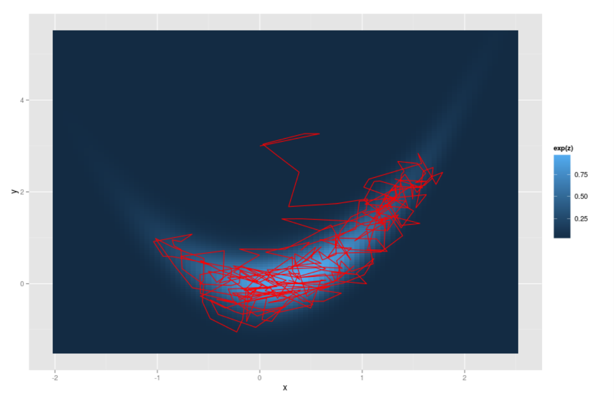
Markov chain Monte Carlo



A random walk:

- start somewhere arbitrary
- make local moves

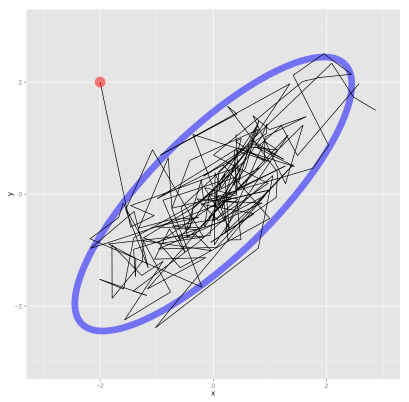
Markov chain Monte Carlo



- 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

Markov chain Monte Carlo

Think of shuffling a pack of cards:

- Can work hard to shuffle perfectly
- Or can make local changes (e.g. cut the deck) to converge (aymptotically) to a perfect shuffle

Markov chain Monte Carlo

Think of shuffling a pack of cards:

- Can work hard to shuffle perfectly
- Or can make local changes (e.g. cut the deck) to converge (aymptotically) to a perfect shuffle
- The goal of MCMC is to find a set of local moves that produce samples (asymtotically) from the right distribution

Markov chain Monte Carlo

Think of shuffling a pack of cards:

- Can work hard to shuffle perfectly
- Or can make local changes (e.g. cut the deck) to converge (aymptotically) to a perfect shuffle
- The goal of MCMC is 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')

Stationary distribution of a Markov chain

A finite state Markov chain with transition matrix T and $X_0 \sim \pi_0$:

$$x_0 \to x_1 \to x_2 \to x_3 \to \cdots \to x_{N-1} \to x_N$$

$x_i \to x_{i+1}$ according to $T(\cdot \to \cdot)$

$$p(x_{i+1} = s_{new} | x_i = s) = T(s \to s_{new})$$

We saw that $X_N \sim T^N \pi_0$

Stationary distribution of a Markov chain

A finite state Markov chain with transition matrix T and $X_0 \sim \pi_0$:

$$x_0 \to x_1 \to x_2 \to x_3 \to \cdots \to x_{N-1} \to x_N$$

$x_i \to x_{i+1}$ according to $T(\cdot \to \cdot)$

$$p(x_{i+1} = s_{new} | x_i = s) = T(s \to s_{new})$$

We saw that $X_N \sim T^N \pi_0$

Perron-Frobenius theorem: always exists distrib. π s. t. $\pi = T\pi$.

π: the stationary distribution of the Markov chain.

Stationary distribution of a Markov chain

A finite state Markov chain with transition matrix T and $X_0 \sim \pi_0$:

$$x_0 \to x_1 \to x_2 \to x_3 \to \cdots \to x_{N-1} \to x_N$$

$x_i \to x_{i+1}$ according to $T(\cdot \to \cdot)$

$$p(x_{i+1} = s_{new} | x_i = s) = T(s \to s_{new})$$

We saw that $X_N \sim T^N \pi_0$

Perron-Frobenius theorem: always exists distrib. π s. t. $\pi = T\pi$.

π: the stationary distribution of the Markov chain.

If $x_0 \sim \pi$, then $X_N \sim \pi$ for all N.

Stationary distribution of a Markov chain

If $x_0 \sim \pi$, then $X_N \sim \pi$ for all N.

$$\mathbb{E}_\pi \left[\frac{1}{N} \sum_{i=1}^{N} g(x_i) \right] = \mathbb{E}_\pi[g]$$

Stationary distribution of a Markov chain

If $x_0 \sim \pi$, then $X_N \sim \pi$ for all N.

$$\mathbb{E}_\pi \left[\frac{1}{N} \sum_{i=1}^{N} g(x_i) \right] = \mathbb{E}_\pi[g]$$

Dependence between x_i's doesn't affect mean.

Stationary distribution of a Markov chain

If $x_0 \sim \pi$, then $X_N \sim \pi$ for all N.

$$\mathbb{E}_\pi \left[\frac{1}{N} \sum_{i=1}^{N} g(x_i) \right] = \mathbb{E}_\pi[g]$$

Dependence between x_i's doesn't affect mean.

MCMC estimate has larger variance (N dependent samples usually has a smaller effective sample size (ESS)).

Is the stationary distribution π unique?

Not always. Example?

Stationary distribution of a Markov chain

Is the stationary distribution π unique?

Not always. Example?

We need our Markov chain to be irreducible:

For each (i, j), there exists an n such that $T_{ij}^n > 0$

Still not sufficient: we need aperiodicity. Example?

Stationary distribution of a Markov chain

Is the stationary distribution π unique?
Not always. Example?

We need our Markov chain to be irreducible:
For each (i, j), there exists an n such that $T^n_{ij} > 0$

Still not sufficient: we need aperiodicity. Example?
Usually ensure aperiodicity by defining a 'lazy' Markov chain.

Ergodicity

A finite-state irreducible aperiodic Markov chain has a unique stationary distribution. For any starting distribution π_0,

$$\pi^N \to \pi \text{ as } N \to \infty$$

$$\frac{1}{N}\sum_{i=1}^{N} g(x_i) \to \mathbb{E}_\pi[g] \qquad \text{(Ergodicity)}$$

Ergodicity

A finite-state irreducible aperiodic Markov chain has a unique stationary distribution. For any starting distribution π_0,

$$\pi^N \to \pi \; as \; N \to \infty$$

$$\frac{1}{N} \sum_{i=1}^{N} g(x_i) \to \mathbb{E}_\pi[g] \qquad \text{(Ergodicity)}$$

A simple algorithm:

- Initialize x_0 from some distribution π_0.
- Run your Markov chain for $(B+N)$ iterations.
- Discard the first B 'burn-in' samples.
- Calculate average using the remaining N samples.

Usually $\mathcal{X}$ is infinite-valued space (e.g. the real line).

$T(x \to \cdot)$ now gives density of next state given current is x

$$P(x_N, x_{N+1}) = \pi(x_N)T(x_{N+1}, x_N)$$

from some distribution π

Markov chains on general state spaces

Usually $\mathcal{X}$ is infinite-valued space (e.g. the real line).

$T(x \to \cdot)$ now gives density of next state given current is x

$$P(x_N, x_{N+1}) = \pi(x_N)T(x_{N+1}, x_N)$$

from some distribution π

π is a stationary distribution if

$$\pi(x) = \int_{\mathcal{X}} \pi(y)T(x, y)\mathrm{d}y$$

Ergodicity: besides irreducibility and aperiodicity, we need 'positive recurrence'.
Informally, the Markov chain should return to any neighbourhood infinitely often.
A harder to establish, but often the case.

Markov chains Monte Carlo

Given $\pi(x) = f(x)/Z$ that is hard to sample from.
Construct a transition kernel T such that

- π is the stationary distribution of T.
- T is irreducible.
- T is aperiodic.
- T is positive recurrent.

Additionally, T should have

- A short burn-in period.
- Fast mixing (large ESS).

Reversible Markov chains

A reversible Markov chain satisfies:

$$\pi(x_N)T(x_{N+1}|x_N) = \pi(x_{N+1})T(x_N|x_{N+1})$$

Also called detailed balance.

Reversible Markov chains

A reversible Markov chain satisfies:

$$\pi(x_N)T(x_{N+1}|x_N) = \pi(x_{N+1})T(x_N|x_{N+1})$$

Also called detailed balance.

Detailed balance implies π is the stationary distribution of T:

$$\begin{aligned}\pi(x_{N+1}) &= \int_{\mathcal{X}} \pi(x_N)T(x_{N+1}|x_N)\mathrm{d}x_N \\ &= \int_{\mathcal{X}} \pi(x_{N+1})T(x_N|x_{N+1})\mathrm{d}x_N = \pi(x_{N+1})\end{aligned}$$

Easy way to verify stationarity or construct T.
Note: converse is not true.

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 to sample from p