Lecture 15: Markov chain Monte Carlo

STAT 545: Introduction to computational statistics

Vinayak Rao

Department of Statistics, Purdue University

November 5, 2018

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

In high dims, hard to find reasonable proposal distributions

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

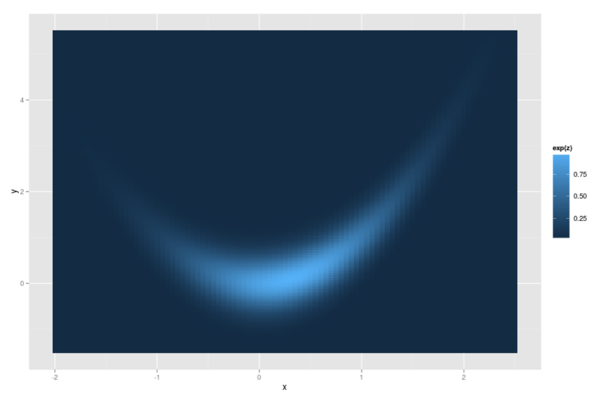
Rather than making independent proposals, exploit previous proposals to make good proposals

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.

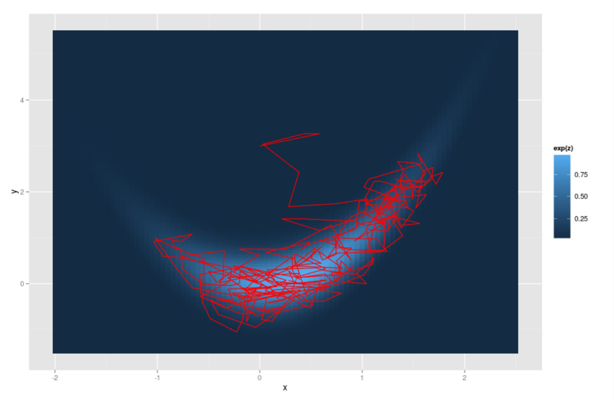
(A Markov chain: future independent of past given present)



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

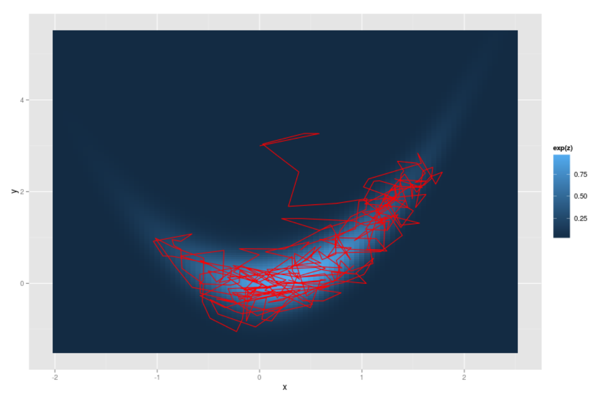
$$p(x,y) \propto \exp\left(-(a-x)^2 - b(y-x^2)^2\right) \quad (\text{here } a = .3, b = 3)$$

Markov chain Monte Carlo



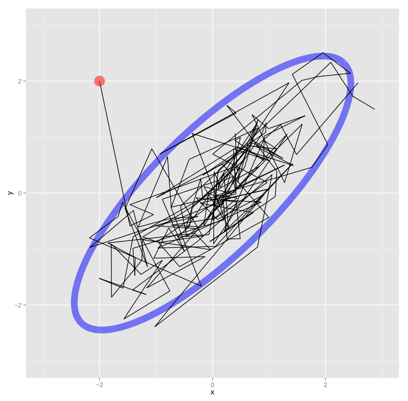
A random walk:

- start somewhere arbitrary
- make local moves



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

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



A random walk over a 2-d Gaussian

Markov chain Monte Carlo

The algorithm at a high level:

- 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.

Need a transition kernel $T(x_{old} \to x_{new})$ that is:

Markov chain Monte Carlo

The algorithm at a high level:

- 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.

Need a transition kernel $T(x_{old} \to x_{new})$ that is:

Correct: The goal of MCMC is to find a set of local moves that produce samples (asymtotically) from the correct distribution

The algorithm at a high level:

- 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.

Need a transition kernel $T(x_{old} \to x_{new})$ that is:

Correct: The goal of MCMC is to find a set of local moves that produce samples (asymtotically) from the correct distribution

Efficient: The art of MCMC is to find inexpensive local moves than coverge rapidly (a chain that 'mixes rapidly')

Ergodicity

What do we mean by correctness?

- For any function h, as $N \to \infty$,

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

- Roughly, MCMC samples look like samples from π

Ergodicity

What do we mean by correctness?

- For any function h, as $N \to \infty$,

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

- Roughly, MCMC samples look like samples from π

What do we mean by efficiency?

- Roughly, MCMC average has similar mean and variation as an average using independent samples from π
- Dependent samples: usually smaller effective sample size

Ergodicity

What do we mean by correctness?

- For any function h, as $N \to \infty$,

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

- Roughly, MCMC samples look like samples from π

What do we mean by efficiency?

- Roughly, MCMC average has similar mean and variation as an average using independent samples from π
- Dependent samples: usually smaller effective sample size

What are conditions for ergodicity?

Ergodicity requires:

Stationarity If x_i is distributed according to π, then so is x_{i+1}

$$\pi(x_{i+1}) = \int_{\mathcal{X}} \pi(x_i) T(x_i \rightarrow x_{i+1}) \mathrm{d}\theta_i$$

Ergodicity requires:

Stationarity If x_i is distributed according to π, then so is x_{i+1}

$$\pi(x_{i+1}) = \int_{\mathcal{X}} \pi(x_i) T(x_i \to x_{i+1}) \mathrm{d}\theta_i$$

Irreducibility Roughly, it should be possible to move between any two parts of space in a finite number of steps

Ergodicity

Ergodicity requires:

Stationarity If x_i is distributed according to π, then so is x_{i+1}

$$\pi(x_{i+1}) = \int_{\mathcal{X}} \pi(x_i) T(x_i \to x_{i+1}) \mathrm{d}\theta_i$$

Irreducibility Roughly, it should be possible to move between any two parts of space in a finite number of steps

Aperiodicity A final technical condition (overcome by allowing self-transitions)

Ergodicity

Ergodicity requires:

Stationarity If x_i is distributed according to π, then so is x_{i+1}

$$\pi(x_{i+1}) = \int_{\mathcal{X}} \pi(x_i) T(x_i \rightarrow x_{i+1}) \mathrm{d}\theta_i$$

Irreducibility Roughly, it should be possible to move between any two parts of space in a finite number of steps

Aperiodicity A final technical condition (overcome by allowing self-transitions)

(Also **positive recurrence** sometimes, but we won't worry too much about this, see slide 14).

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)).

Stationary distribution of a Markov chain

Does a stationary distribution always exist?

For a finite-state chain, yes (Perron-Frobenius theorem).

For the general case, no. Counterexample?

Stationary distribution of a Markov chain

Does a stationary distribution always exist?

For a finite-state chain, yes (Perron-Frobenius theorem).

For the general case, no. Counterexample?

Is the stationary distribution π unique?

Not always. Example?

Stationary distribution of a Markov chain

Does a stationary distribution always exist?

For a finite-state chain, yes (Perron-Frobenius theorem).

For the general case, no. Counterexample?

Is the stationary distribution π unique?

Not always. Example?

We need our Markov chain to be irreducible:

For each (x_{old}, x_{new}), there exists an n such that

$$T^n(x_{old} \to x_{new}) > 0$$

Still not sufficient: we need aperiodicity. Example?

Stationary distribution of a Markov chain

Does a stationary distribution always exist?
For a finite-state chain, yes (Perron-Frobenius theorem).
For the general case, no. Counterexample?

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

We need our Markov chain to be irreducible:
For each (x_{old}, x_{new}), there exists an n such that

$$T^n(x_{old} \to x_{new}) > 0$$

Still not sufficient: we need aperiodicity. Example?
Usually ensure aperiodicity by defining a 'lazy' Markov chain that can have self-transitions.

Ergodicity

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

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

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

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

Now ergodicity also needs ‘positive recurrence’.
Informally, the Markov chain should return to any neighborhood infinitely often.
Harder to establish, but often the case.

Checking stationarity is not easy. Common to use **reversibility**

Reversible Markov chains

Checking stationarity is not easy. Common to use **reversibility**

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

Checking stationarity is not easy. Common to use **reversibility**

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

(just integrate both sides w.r.t. x_N)

Reversible Markov chains

Checking stationarity is not easy. Common to use **reversibility**

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

(just integrate both sides w.r.t. x_N)

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

MCMC: A FIRST LOOK

Find a transition function $T(\cdot \to \cdot)$ with stationary distrib. 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

Find a transition function $T(\cdot \to \cdot)$ with stationary distrib. 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

Find a transition function $T(\cdot \to \cdot)$ with stationary distrib. 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