Lecture 9: Markov Chain Monte Carlo

Jacob Steinhardt

September 24, 2020

Last Time

- Rejection sampling
- Markov chain review

Last Time

- Rejection sampling
- Markov chain review

This time: Markov chain Monte Carlo

- Gibbs sampling
- Metropolis-Hastings

Gibbs Sampling: Motivation

- Have an arbitrary distribution $p(x_1, \ldots, x_n)$ that we want to sample from

Gibbs Sampling: Motivation

- Have an arbitrary distribution $p(x_1, \ldots, x_n)$ that we want to sample from
- Current tool: rejection sampling
  - Proposal distribution $q(x_1, \ldots, x_n)$ for all x_i at once
  - Issue: too slow (typically exponentially small acceptance rate in n)
  - E.g. even if x_i are independent, and $q(x_i)/p(x_i) \leq 1.1$, need 1.1^n tries ($\approx 2.5 \cdot 10^{41}$ for $n = 1000$)

Gibbs Sampling: Motivation

- Have an arbitrary distribution $p(x_1, \ldots, x_n)$ that we want to sample from
- Current tool: rejection sampling
  - Proposal distribution $q(x_1, \ldots, x_n)$ for all x_i at once
  - Issue: too slow (typically exponentially small acceptance rate in n)
  - E.g. even if x_i are independent, and $q(x_i)/p(x_i) \leq 1.1$, need 1.1^n tries ($\approx 2.5 \cdot 10^{41}$ for $n = 1000$)
- Idea behind Gibbs sampling: change one variable at a time (Markov chain)

Gibbs Sampling: Algorithm

Algorithm:

- Initialize $(x_1, \dots, x_n)$ arbitrarily
- Repeat:
  - Pick i (randomly or sequentially)
  - Re-sample x_i from $p(x_i \mid x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$ (often denote $p(x_i \mid x_{-i})$)

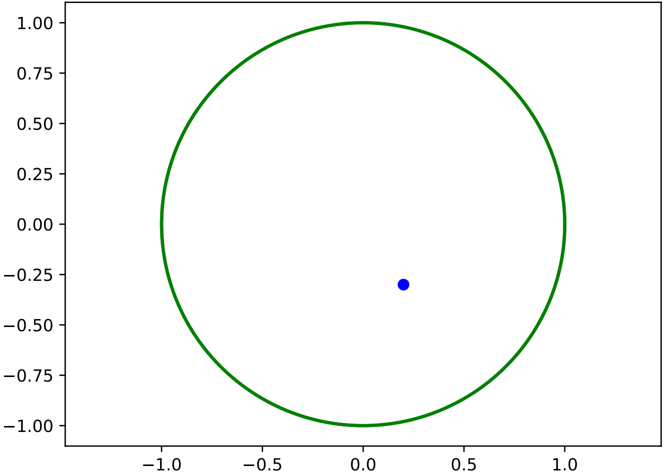
Gibbs Sampling: Algorithm

Algorithm:

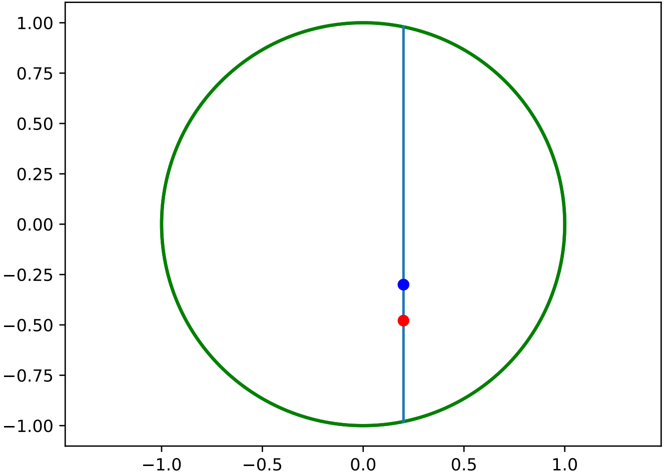
- Initialize $(x_1, \ldots, x_n)$ arbitrarily
- Repeat:
  - Pick i (randomly or sequentially)
  - Re-sample x_i from $p(x_i \mid x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ (often denote $p(x_i \mid x_{-i})$)

Defines a Markov chain, and can prove that the stationary distribution is $p(x_1, \ldots, x_n)$ (!!).

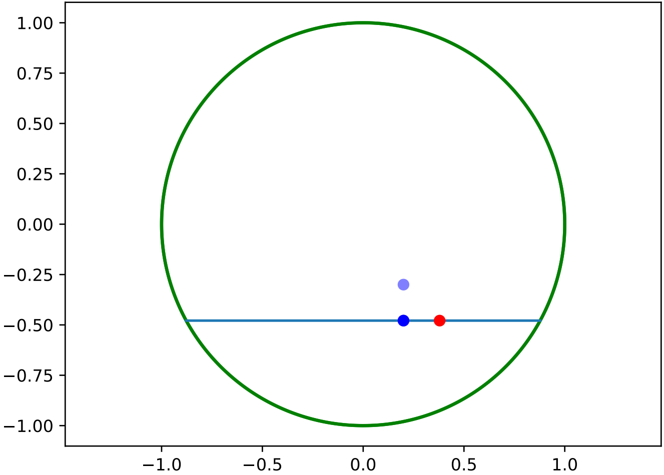
Gibbs Sampling: Unit Circle Example



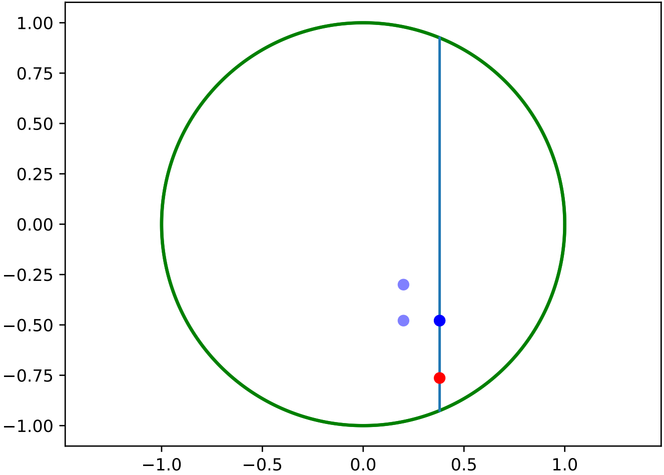
Gibbs Sampling: Unit Circle Example



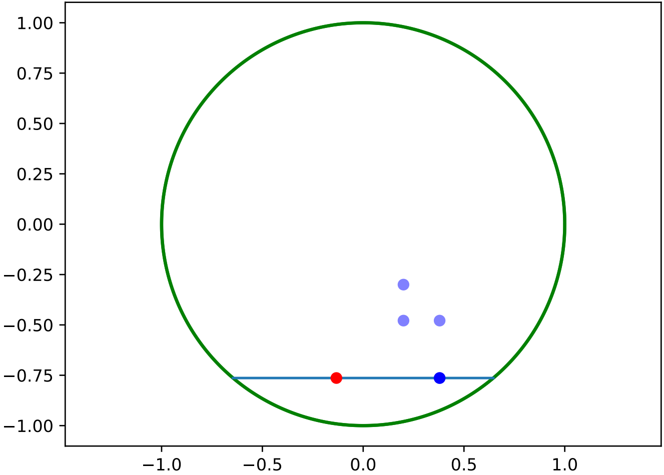
Gibbs Sampling: Unit Circle Example



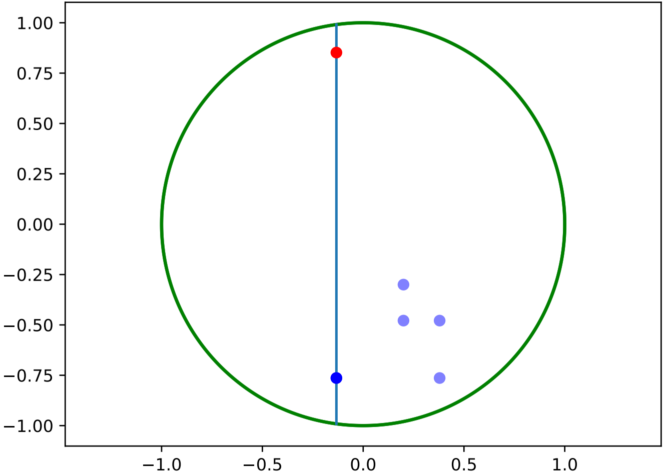
Gibbs Sampling: Unit Circle Example



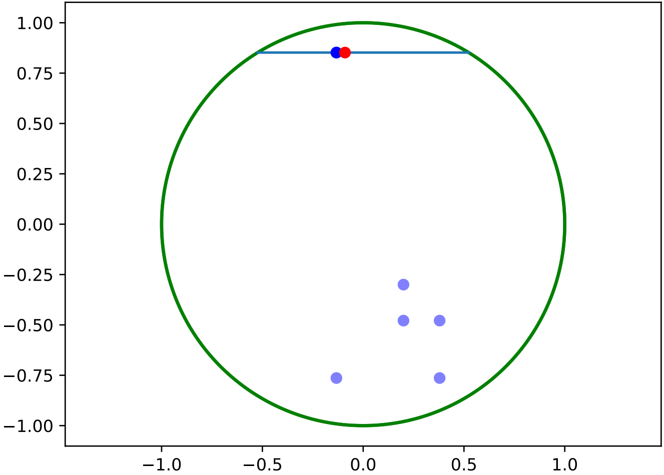
Gibbs Sampling: Unit Circle Example



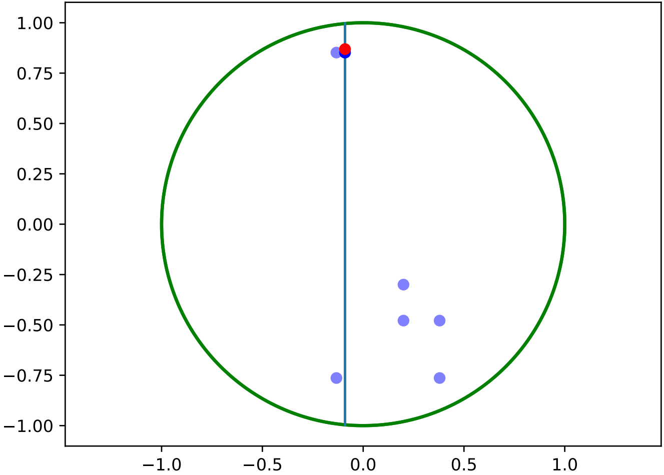
Gibbs Sampling: Unit Circle Example



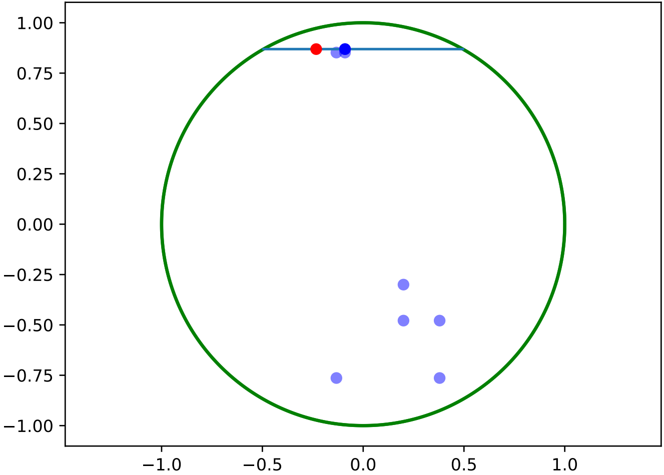
Gibbs Sampling: Unit Circle Example



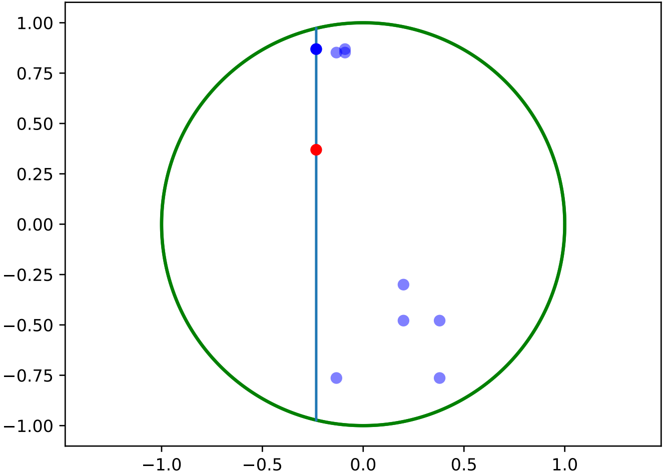
Gibbs Sampling: Unit Circle Example



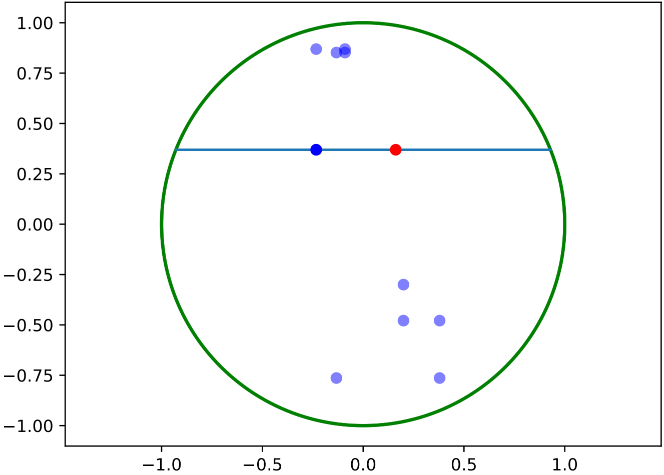
Gibbs Sampling: Unit Circle Example



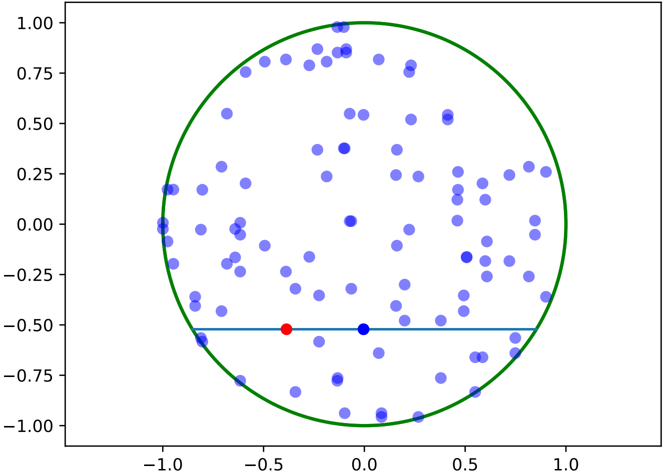
Gibbs Sampling: Unit Circle Example



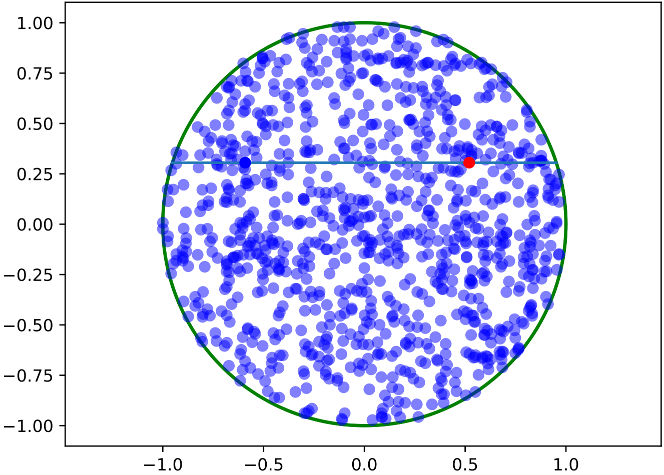
Gibbs Sampling: Unit Circle Example



Gibbs Sampling: Unit Circle Example

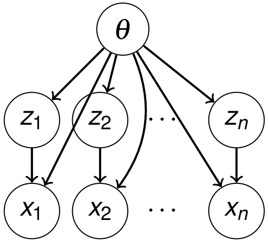


Gibbs Sampling: Unit Circle Example



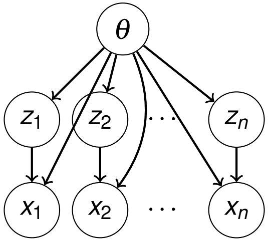
Gibbs Sampling for Hierarchical Models

Recall hierarchical models (e.g. height and gender example)



Gibbs Sampling for Hierarchical Models

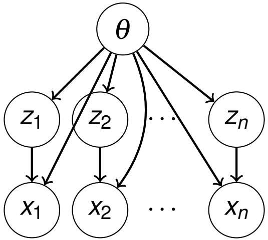
Recall hierarchical models (e.g. height and gender example)



Suppose we want to do Gibbs sampling for this model

Gibbs Sampling for Hierarchical Models

Recall hierarchical models (e.g. height and gender example)

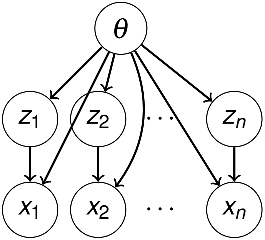


Suppose we want to do Gibbs sampling for this model

- Sample z_i: $p(z_i \mid x_i, \theta) \propto \underbrace{p(z_i \mid \theta)}_{\text{prior}} \underbrace{p(x_i \mid z_i)}_{\text{likelihood}}$

Gibbs Sampling for Hierarchical Models

Recall hierarchical models (e.g. height and gender example)



Suppose we want to do Gibbs sampling for this model

- Sample z_i: $p(z_i \mid x_i, \theta) \propto \underbrace{p(z_i \mid \theta)}_{\text{prior}} \underbrace{p(x_i \mid z_i)}_{\text{likelihood}}$
- Sample θ (e.g. μ_0 for height/gender model):

$$p(\mu_0 \mid z_{1:n}, x_{1:n}) \propto \underbrace{p(\mu_0)}_{\text{prior}} \cdot \underbrace{\prod_{i:z_i=0} \exp(-(x_i - \mu_0)^2/2\sigma^2)}_{\text{likelihood}}$$

Proof of stationary distribution

Assuming chain is ergodic, just need to show stationary distribution is preserved.

Suppose $x \sim p$ and x' is obtained from x by Gibbs sampling update. Want to show that x' is also distributed according to p.

If index i is updated, then $x' = (x_1, \ldots, x_{i-1}, x'_i, x_{i+1}, \ldots)$, where $x'_i \sim p(x_i \mid x_1, \ldots, x_{i-1}, x_{i+1}, \ldots)$.

Indices $\neq i$ distributed according to p, and $x'_i \mid x'_{\neq i}$ is as well, so x' follows p.

Ergodicity: counterexample

Suppose that $x_1, x_2 \in \{0, 1\}$ with following probability table:

	0	1
0	0.5	0.0
1	0.0	0.5

What will Gibbs sampling do?

Gibbs Sampling: Summary

- Repeatedly sample from $p(x_i \mid x_{-i})$
- Creates Markov chain whose stationary distribution is $p(x_1, \dots, x_n)$
- Flexible: conditional $p(x_i \mid x_{-i})$ one-dimensional, easy to sample from
- Don't need to "get lucky" with graphical model structure
- Extensions, e.g. block Gibbs sampling

Metropolis-Hastings: Idea

- Gibbs sampling: one possible Markov chain

Metropolis-Hastings: Idea

- Gibbs sampling: one possible Markov chain
- Is there a more general strategy?

Metropolis-Hastings: Idea

- Gibbs sampling: one possible Markov chain
- Is there a more general strategy?
- Yes! Combine with idea of rejection sampling

Metropolis-Hastings: Idea

- Gibbs sampling: one possible Markov chain
- Is there a more general strategy?
- Yes! Combine with idea of rejection sampling
- Given any “proposed Markov chain” $q(x^{\text{new}} \mid x^{\text{old}})$, will combine with an accept/reject step to create new Markov chain with the correct stationary distribution

Metropolis-Hastings: Algorithm

Proposal distribution: $q(x^{\text{new}} \mid x^{\text{old}})$

Given x^{old}:

- Sample x^{new} from q
- With probability [ ], accept (replace x^{old} with x^{new})
- Otherwise, reject (keep x^{old})

Metropolis-Hastings: Algorithm

Proposal distribution: $q(x^{\text{new}} \mid x^{\text{old}})$

Given x^{old}:

- Sample x^{new} from q
- With probability $\boxed{\qquad \frac{p(x^{\text{new}})}{p(x^{\text{old}})} \qquad}$, accept (replace x^{old} with x^{new})
- Otherwise, reject (keep x^{old})

Metropolis-Hastings: Algorithm

Proposal distribution: $q(x^{\text{new}} \mid x^{\text{old}})$

Given x^{old}:

- Sample x^{new} from q
- With probability $\boxed{\min\left(1, \frac{p(x^{\text{new}})}{p(x^{\text{old}})} \frac{q(x^{\text{old}}|x^{\text{new}})}{q(x^{\text{new}}|x^{\text{old}})}\right)}$, accept (replace x^{old} with x^{new})
- Otherwise, reject (keep x^{old})

Metropolis-Hastings: Algorithm

Proposal distribution: $q(x^{\text{new}} \mid x^{\text{old}})$

Given x^{old}:

- Sample x^{new} from q
- With probability $\boxed{\min\left(1, \frac{p(x^{\text{new}})}{p(x^{\text{old}})} \frac{q(x^{\text{old}} | x^{\text{new}})}{q(x^{\text{new}} | x^{\text{old}})}\right)}$, accept (replace x^{old} with x^{new})
- Otherwise, reject (keep x^{old})

Metropolis-Hastings: Algorithm

Proposal distribution: $q(x^{\text{new}} \mid x^{\text{old}})$

Given x^{old}:

- Sample x^{new} from q
- With probability $\boxed{\min\left(1, \frac{p(x^{\text{new}})}{p(x^{\text{old}})} \frac{q(x^{\text{old}} | x^{\text{new}})}{q(x^{\text{new}} | x^{\text{old}})}\right)}$, accept (replace x^{old} with x^{new})
- Otherwise, reject (keep x^{old})

Gibbs sampling: special choice of q where we always accept!

Proof sketch: Detailed balance

Can show that if an ergodic Markov chain satisfies
$\bar{p}(x)A(x' \mid x) = \bar{p}(x')A(x \mid x')$ for all x, x', then it has stationary distribution $\bar{p}$.

This condition is called **detailed balance**.

Metropolis-Hastings sets probabilities so that detailed balance holds.

Mixing time

Performance of MCMC algorithms governed by **mixing time**: how long it takes to get close to stationary distribution.

Mixing time can vary dramatically, from nearly linear to exponential in number of variables.

[mixing time examples: on board]