Lecture 11: Markov Chain Monte Carlo

Jacob Steinhardt

February 24, 2020

Announcements

- Jacob away Wed-Fri (no office hours)
- Lecture 12: Guest lecture (Clara Wong-Fannjiang)
- HW2 due, HW3 released
- Moritz back next week!

Last Time

- Rejection sampling
- Importance sampling

Last Time

- Rejection sampling
- Importance sampling

This time: Markov chain Monte Carlo

- Markov chain review
- Gibbs sampling
- Metropolis-Hastings

Review: Markov Chains

Markov chain: sequence $x_1, x_2, \ldots, x_T$ where distribution of x_t depends only on x_{t-1}

Defined by *transition distribution* $A(x^{\text{new}} \mid x^{\text{old}})$, together with initial state x_1

Examples:

- Random walk on a graph
- Repeatedly shuffling a deck of cards
- Process defined by

$$x_1 = 0, \quad x_t \mid x_{t-1} \sim N(0.9x_{t-1}, 1)$$

Markov Chains: Stationary Distribution

All "nice enough" Markov chains have the property that if T is large enough, the distribution over x_T is almost independent of x_1, and converges to some distribution $\bar{p}(x)$ as $T \to \infty$.

Markov Chains: Stationary Distribution

All "nice enough" Markov chains have the property that if T is large enough, the distribution over x_T is almost independent of x_1, and converges to some distribution $\bar{p}(x)$ as $T \to \infty$.

$\bar{p}(x)$ is called the *stationary distribution*, and the technical condition for "nice enough" is that the Markov chain is *ergodic*.

Markov Chains: Stationary Distribution

All “nice enough” Markov chains have the property that if T is large enough, the distribution over x_T is almost independent of x_1, and converges to some distribution $\bar{p}(x)$ as $T \to \infty$.

$\bar{p}(x)$ is called the *stationary distribution*, and the technical condition for “nice enough” is that the Markov chain is *ergodic*.

The distribution $\bar{p}(x)$ is also what we get if we count how many times x_t visits each state, as $T \to \infty$.

Markov Chains: Mixing Time

The *mixing time* is how long it takes for x_T to be close to the stationary distribution (we won't define this formally).

Markov Chains: Mixing Time

The *mixing time* is how long it takes for x_T to be close to the stationary distribution (we won't define this formally).

Example: card shuffling

- Mixing time is how many shuffles we need for deck to be "almost random"

Markov Chains: Mixing Time

The *mixing time* is how long it takes for x_T to be close to the stationary distribution (we won't define this formally).

Example: card shuffling

- Mixing time is how many shuffles we need for deck to be "almost random"

Other examples:

- Random walk on complete graph with n vertices
- Random walk on path of length n

The Annals of Applied Probability
1992, Vol. 2, No. 2, 294–313

TRAILING THE DOVETAIL SHUFFLE TO ITS LAIR

BY DAVE BAYER[1] AND PERSI DIACONIS[2]

Columbia University and Harvard University

We analyze the most commonly used method for shuffling cards. The main result is a simple expression for the chance of any arrangement after any number of shuffles. This is used to give sharp bounds on the approach to randomness: $\frac{3}{2}\log_2 n + \theta$ shuffles are necessary and sufficient to mix up n cards.

Key ingredients are the analysis of a card trick and the determination of the idempotents of a natural commutative subalgebra in the symmetric group algebra.

1. Introduction. The dovetail, or riffle shuffle is the most commonly used method of shuffling cards. Roughly, a deck of cards is cut about in half and then the two halves are riffled together. Figure 1 gives an example of a riffle shuffle for a deck of 13 cards.

A mathematically precise model of shuffling was introduced by Gilbert and Shannon [see Gilbert (1955)] and independently by Reeds (1981). A deck of n cards is cut into two portions according to a binomial distribution; thus, the chance that k cards are cut off is $\binom{n}{k}/2^n$ for $0 \le k \le n$. The two packets are then riffled together in such a way that cards drop from the left or right heaps

Markov chains: recap

- Governed by proposal distribution $A(x^{\mathrm{new}} \mid x^{\mathrm{old}})$
- Stationary distribution: limiting distribution of x_T
- Mixing time: how long it takes to get to stationary distribution

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 $p(x_i \mid x_{-i})$)

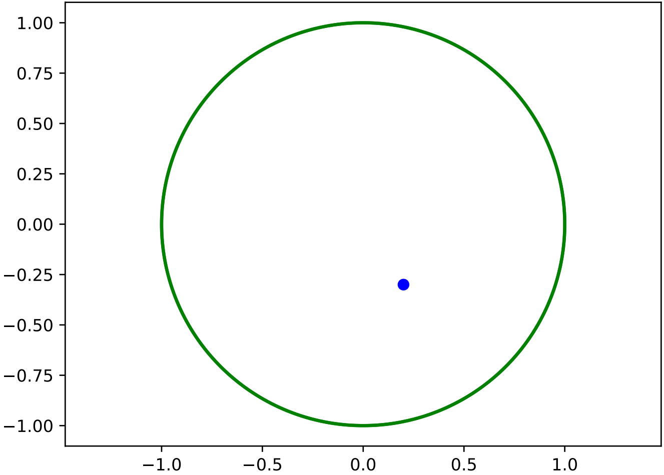
Gibbs Sampling: Algorithm

Algorithm:

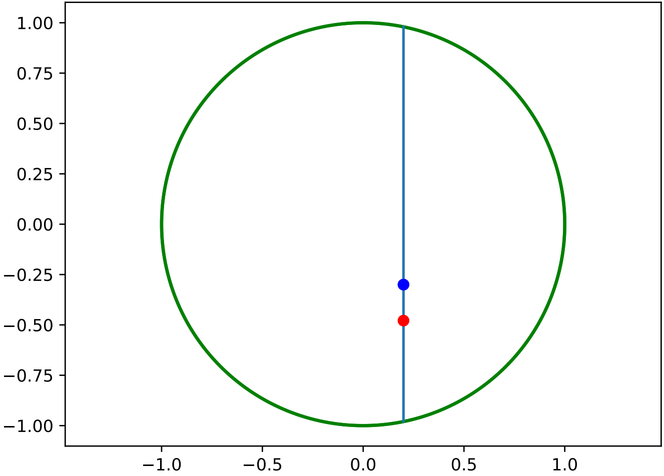
- Initialize $(x_1, \ldots, x_n)$ arbitrarily
- Repeat:
  - Pick i (randomly or sequentially)
  - Re-sample x_i from $p(x_i \mid x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ (often denote $p(x_i \mid x_{-i})$)

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

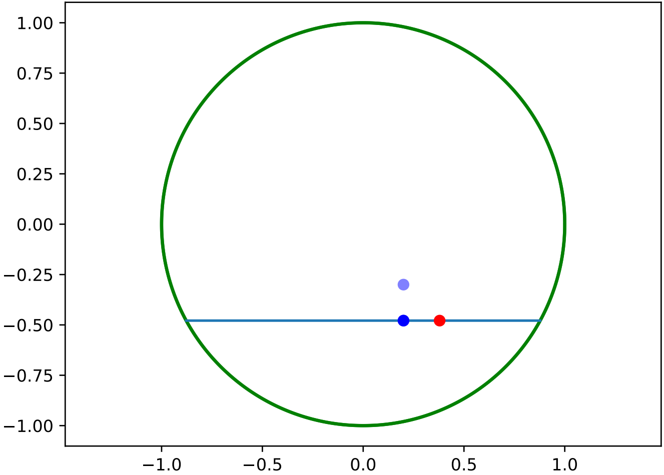
Gibbs Sampling: Unit Circle Example



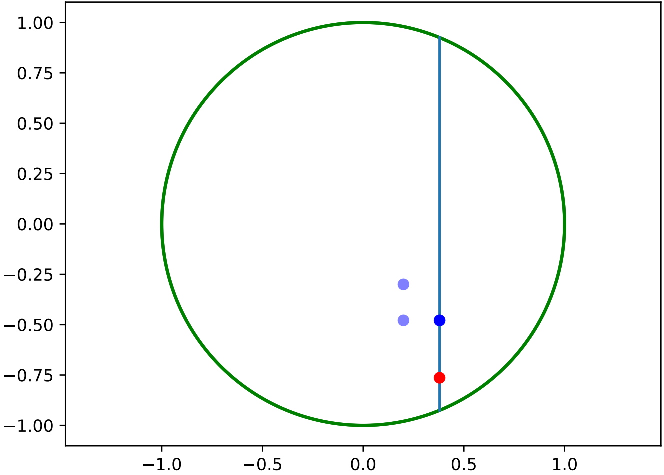
Gibbs Sampling: Unit Circle Example



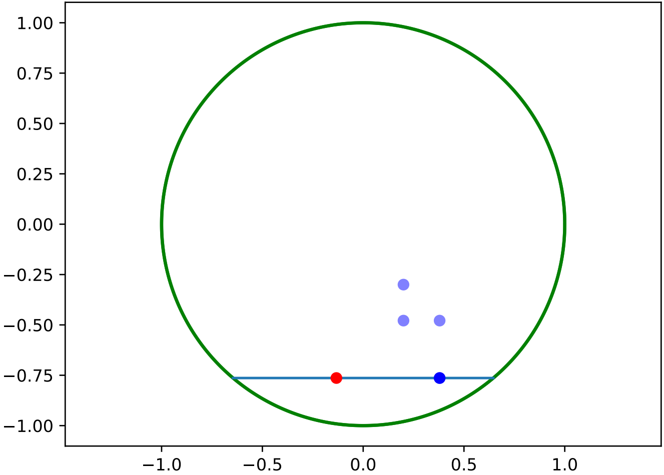
Gibbs Sampling: Unit Circle Example



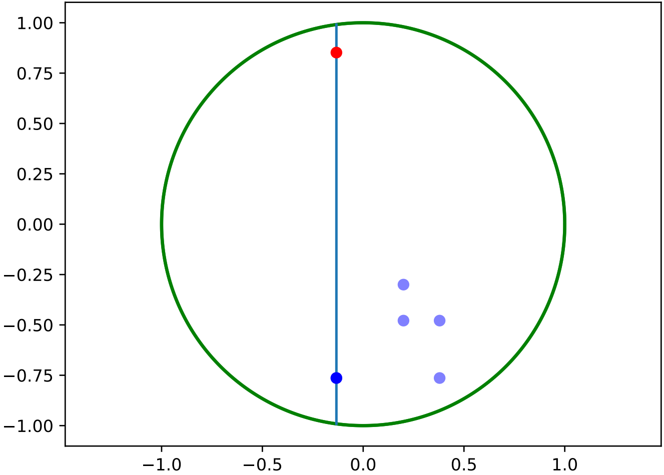
Gibbs Sampling: Unit Circle Example



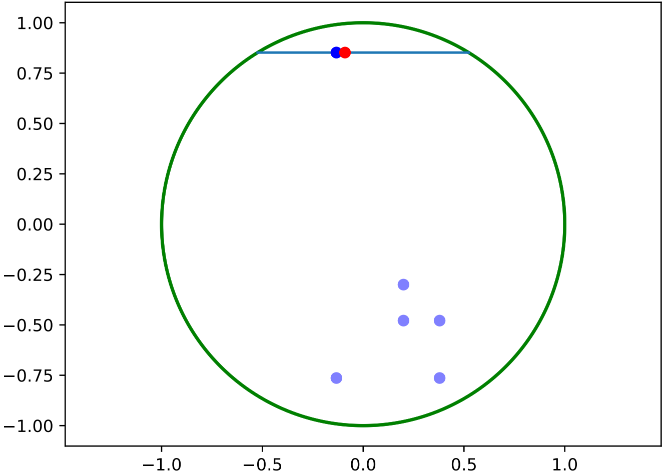
Gibbs Sampling: Unit Circle Example



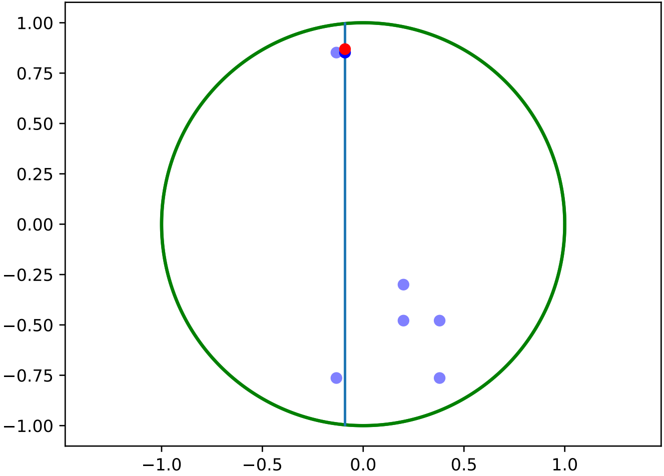
Gibbs Sampling: Unit Circle Example



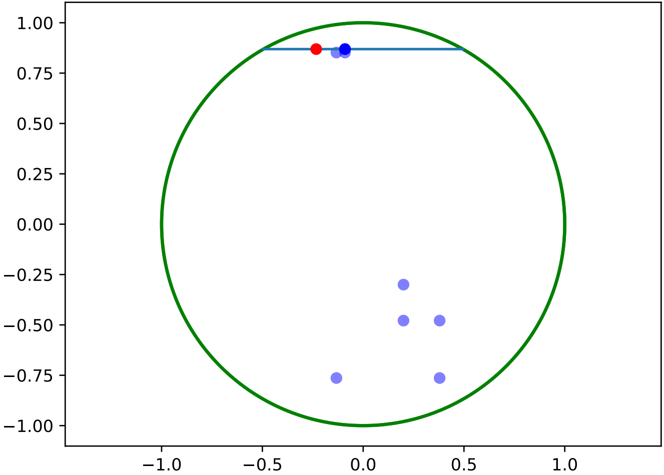
Gibbs Sampling: Unit Circle Example



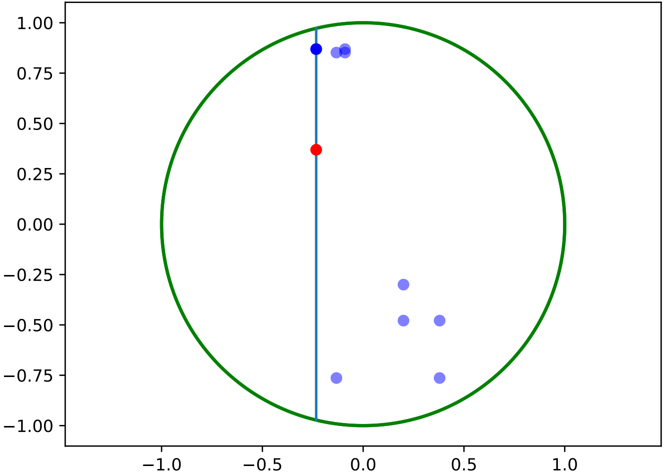
Gibbs Sampling: Unit Circle Example



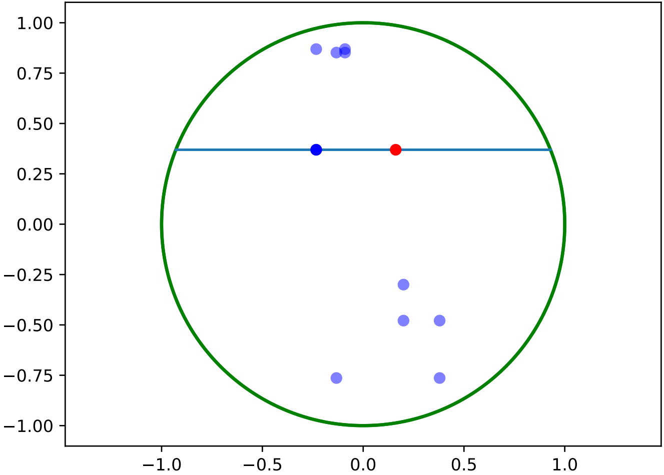
Gibbs Sampling: Unit Circle Example



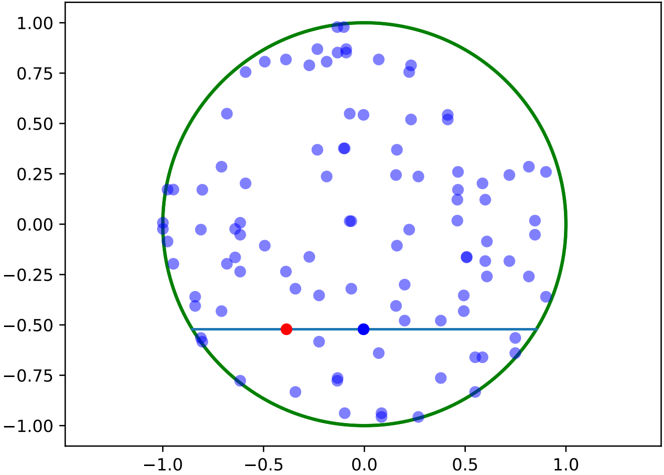
Gibbs Sampling: Unit Circle Example



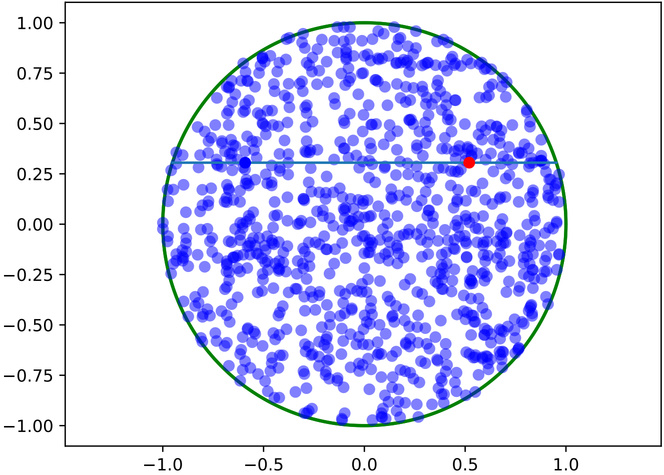
Gibbs Sampling: Unit Circle Example



Gibbs Sampling: Unit Circle Example

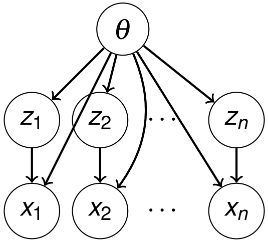


Gibbs Sampling: Unit Circle Example



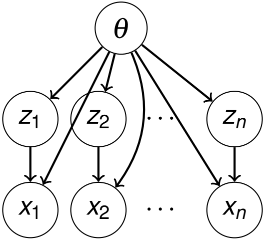
Gibbs Sampling for Hierarchical Models

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



Gibbs Sampling for Hierarchical Models

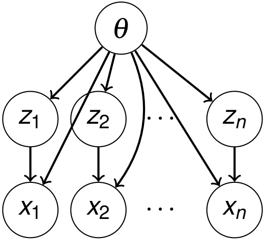
Recall hierarchical models (e.g. height and gender example)



Suppose we want to do Gibbs instead of EM

Gibbs Sampling for Hierarchical Models

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

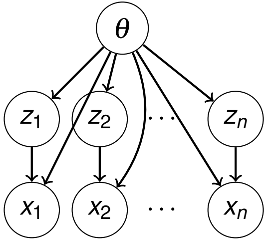


Suppose we want to do Gibbs instead of EM

- 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 instead of EM

- 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):

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

Gibbs Sampling: Summary

- Repeatedly sample from $p(x_i \mid x_{-i})$
- Creates Markov chain whose stationary distribution is $p(x_1, \ldots, x_n)$
- 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 $\boxed{}$, 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 $\boxed{\frac{p(x^{\text{new}})}{p(x^{\text{old}})}}$, 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 $\boxed{\quad \frac{p(x^{\text{new}})}{p(x^{\text{old}})} \frac{q(x^{\text{old}}|x^{\text{new}})}{q(x^{\text{new}}|x^{\text{old}})}}$, 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!