Introduction to Machine Learning CMU-10701

Markov Chain Monte Carlo Methods

Barnabás Póczos & Aarti Singh

Contents

- Markov Chain Monte Carlo Methods
  - Goal & Motivation
- Sampling
  - Rejection
  - Importance
- Markov Chains
  - Properties
- MCMC sampling
  - Hastings-Metropolis
  - Gibbs

Monte Carlo Methods

The importance of MCMC

- A recent survey places the **Metropolis algorithm** among the

  10 algorithms that have had the *greatest influence* on the development and practice of science and engineering in the 20th century (Beichl&Sullivan, 2000).

- The Metropolis algorithm is an instance of a large class of sampling algorithms, known as **Markov chain Monte Carlo** (MCMC).

MCMC Applications

MCMC plays significant role in **statistics, econometrics, physics and computing science**.

- **Sampling from high-dimensional, complicated distributions**
- **Bayesian inference and learning**
  - *Marginalization* $p(x) = \int_Z p(x, z) dz$
  - *Normalization* $p(x|y) = \frac{p(y|x)p(x)}{\int_X p(y|x)p(x)dx}$
  - *Expectation* $\mathbb{E}_{p(x)}(f(x)) = \int_X f(x)p(x)\, dx$
- **Global optimization** $\arg\max_x f(x)$

The Monte Carlo principle

One "tiny" problem…

- Monte Carlo methods need sample from distribution $p(x)$.
- When $p(x)$ has standard form, e.g. Uniform or Gaussian, it is straightforward to sample from it using easily available routines.
- However, when this is not the case, we need to introduce more sophisticated sampling techniques. $\Rightarrow$ MCMC sampling

Sampling

- Rejection sampling
- Importance sampling

Main Goal

***Sample from distribution $p(x)$* that is only known up to a proportionality constant**

For example,

$$p(x) \propto 0.3 \exp(-0.2x^2) + 0.7 \exp(-0.2(x - 10)^2)$$

Rejection Sampling

Rejection Sampling Conditions

Suppose that

- $p(x)$ is known up to a proportionality constant

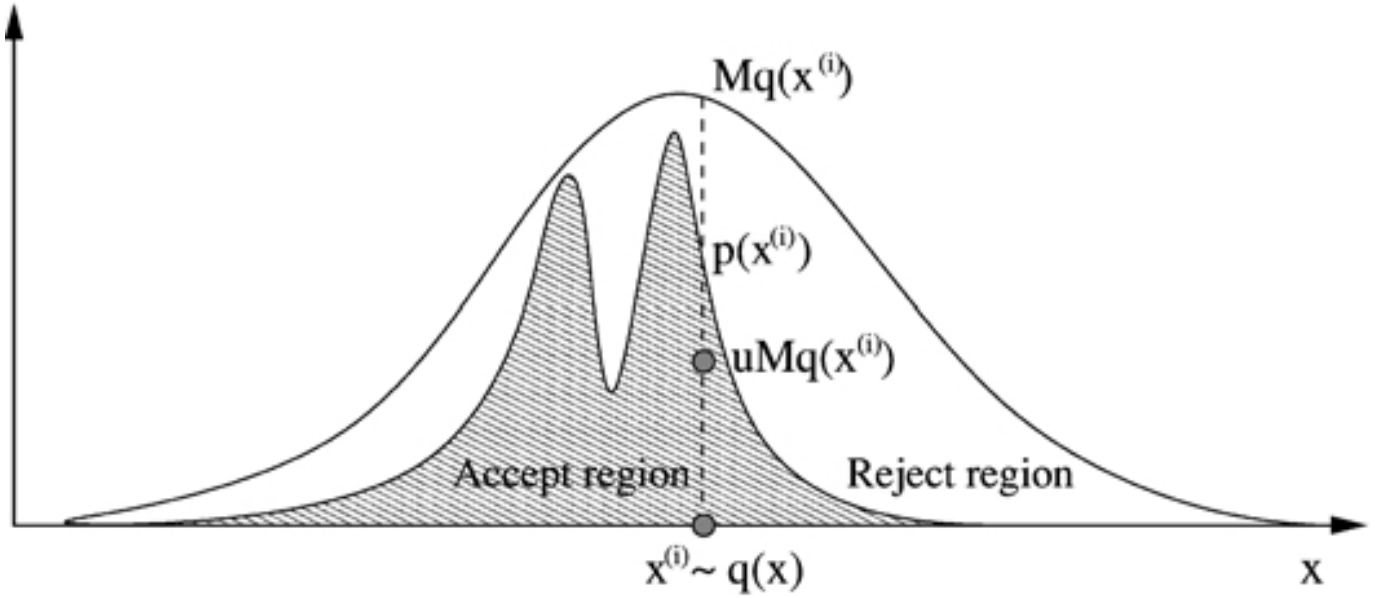
$$p(x) \propto 0.3 \exp(-0.2x^2) + 0.7 \exp(-0.2(x - 10)^2)$$

- It is easy to sample from $q(x)$ that satisfies $p(x) \leq M\, q(x)$, $M < \infty$

- M is known

Rejection Sampling Algorithm

1) $i = 1$
2) Repat until $i = N$
   a) Sample: $x^{(i)} \sim q(x)$ and $u \sim U_{[0,1]}$
   b) If $u < \frac{p(x^{(i)})}{Mq(x^{(i)})}$ then accept $x^{(i)}$ and $i = i + 1$,
   otherwise reject $x^{(i)}$.



Rejection Sampling

Theorem

The accepted $x^{(i)}$ can be shown to be sampled with probability $p(x)$ (Robert & Casella, 1999, p. 49).

Severe limitations:

- It is not always possible to bound p(x)/q(x) with a reasonable constant M over the whole space X.
- If M is too large, the acceptance probability is too small.
- In high dimensional spaces it can be exponentially slow to sample points. (The points usually will be rejected)

Importance Sampling

Importance Sampling

Goal: Sample from distribution p(x) that is only known up to a proportionality constant

- Importance sampling is an alternative "classical" solution that goes back to the 1940's.

- Let us introduce, again, an arbitrary importance proposal distribution *q(x)* such that its support includes the support of *p(x)*.

- Then we can rewrite *I(f)* as follows:

$$
\begin{aligned}
I(f) &= \int f(x)p(x)dx \\
&= \int f(x)\frac{p(x)}{q(x)}q(x)dx \\
&= \int f(x)w(x)q(x)dx
\end{aligned}
$$

Importance Sampling

$$
\begin{aligned}
I(f) &= \int f(x)p(x)dx \\
&= \int f(x)\frac{p(x)}{q(x)}q(x)dx \qquad\qquad w(x) = \frac{p(x)}{q(x)} \\
&= \int f(x)w(x)q(x)dx
\end{aligned}
$$

Consequently,

- ⋆ if one can draw N i.i.d. $x^{(i)}$ $i = 1, \ldots, N$ from $q(x)$,
- ⋆ and evaluate $w(x^{(i)})$, then

$\Rightarrow$ possible Monte Carlo estimate of $I(f)$ is

$$\widehat{I}_N(f) = \frac{1}{N}\sum_{i=1}^{N} f(x^{(i)})w(x^{(i)})$$

Importance Sampling

$$\hat{I}_N(f) = \frac{1}{N} \sum_{i=1}^{N} f(x^{(i)}) w(x^{(i)})$$

Theorem

- This estimator is unbiased
- Under weak assumptions, the strong law of large numbers applies:

$$\hat{I}_N(f) = \xrightarrow[N \to \infty]{a.s.} I(f) = \int_{\chi} f(x) p(x) dx$$

Some proposal distributions $q(x)$ will obviously be preferable to others.

Which one should we choose?

Importance Sampling

$$\hat{I}_N(f) = \frac{1}{N} \sum_{i=1}^{N} f(x^{(i)}) w(x^{(i)})$$

Theorem

- This estimator is unbiased
- Under weak assumptions, the strong law of large numbers applies:

$$\hat{I}_N(f) = \xrightarrow[N \to \infty]{a.s.} I(f) = \int_{\chi} f(x) p(x) dx$$

Some proposal distributions $q(x)$ will obviously be preferable to others.

$$Var_{q(x)}[\hat{I}_N(f)] = \mathbb{E}_{q(x)}[f^2(x) w^2(x) - I^2(f)]$$

Importance Sampling

Find one that minimizes the variance of the estimator!

$$Var_{q(x)}[\hat{I}_N(f)] = \mathbb{E}_{q(x)}[f^2(x)w^2(x) - I^2(f)]$$

Theorem

The variance is minimal when we adopt the following ***optimal importance distribution:***

$$q^*(x) = \frac{|f(x)|p(x)}{\int |f(x)|p(x)dx}$$

Importance Sampling

- The optimal proposal is not very useful in the sense that it is not easy to sample from

$$q^*(x) = \frac{|f(x)|p(x)}{\int |f(x)|p(x)dx}$$

- High sampling efficiency is achieved when we focus on sampling from p(x) in the important regions where |f (x)|p(x) is relatively large; hence the name ***importance sampling***

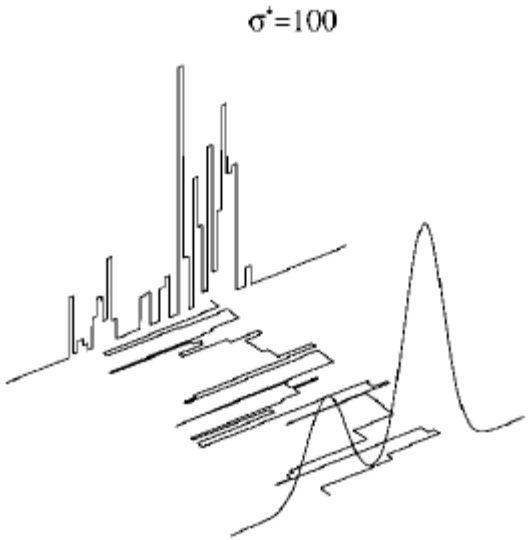
- Importance sampling estimates can be **super-efficient**:

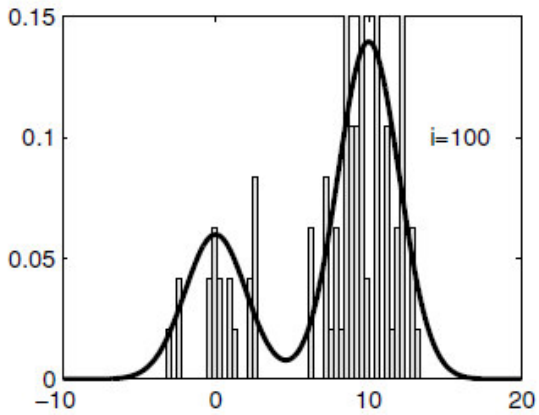
  For a given function $f(x)$, it is possible to find a distribution $q(x)$ that yields an estimate with a lower variance than when using $q(x) = p(x)$!

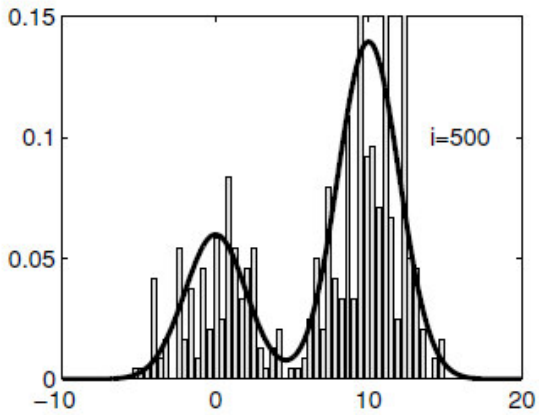
- In high dimensions it is not efficient either…

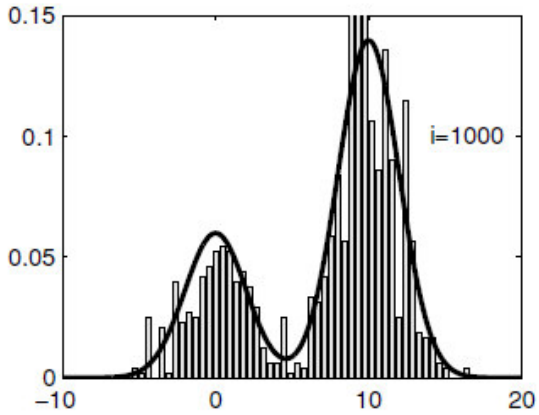
MCMC sampling - Main ideas

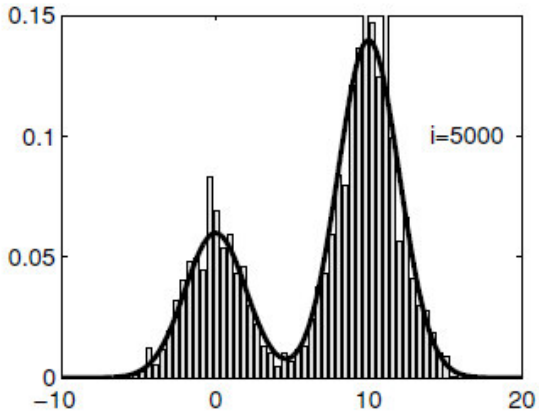
Create a Markov chain, which has the desired limiting distribution!











Markov Chains, stationary distribution

Definition:

[stationary distribution, invariant distribution, steady state distributions]

The distribution $\pi = (\pi_1, \ldots, \pi_k)$ is **stationary** distribution if $\pi_i \geq 0 \ \forall i$, $\sum_{i=1}^{T} \pi_i = 1$, and $\pi \mathbf{T} = \pi$.

T maintians π. If we start the chain from distribution π, in the next step the distribution remains the same.

$$(0.22, 0.41, 0.37) \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0.1 & 0.9 \\ 0.6 & 0.4 & 0 \end{bmatrix} = (0.22, 0.41, 0.37)$$

The stationary distribution might be not unique (e.g. T= identity matrix)

Markov Chains, limit distributions

Some Markov chains have **unique limit distribution:**

$$T = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0.1 & 0.9 \\ 0.6 & 0.4 & 0 \end{bmatrix}$$

If the probability vector for the initial state is $\mu(x^{(1)}) = (0.5, 0.2, 0.3)$

it follows that $\mu(x^{(1)})T = (0.2, 0.6, 0.2)$

and, after several iterations (multiplications by T)

$\mu(x^{(1)})T^t \to p(x) = (0.22, 0.41, 0.37)$ **limit distribution**

no matter what initial distribution $\mu(x^1)$ was.

$$T^\infty = \begin{bmatrix} 0.22 & 0.41 & 0.37 \\ 0.22 & 0.41 & 0.37 \\ 0.22 & 0.41 & 0.37 \end{bmatrix}$$

The chain has forgotten its past.

Markov Chains

Our goal is to **find conditions** under which the Markov chain

converges to a unique limit distribution (independently from its starting state distribution)

Observation:

If this limiting distribution exists, it has to be the stationary distribution.

Limit Theorem of Markov Chains

Theorem:

If the Markov chain is **Irreducible** and **Aperiodic**, then:

1. $\exists! \pi = (\pi_1, \ldots \pi_n)$ stationary distribution

2. $\lim_{t \to \infty} \frac{E(\textit{ number of chain visits state i in t steps})}{t} = \pi_i$

3. $\lim_{t \to \infty} \Pr(X_n = i) = \pi_i \ \forall i$

4. $\lim_{t \to \infty} \mathbf{v}\mathbf{T}^t = \pi \ \forall \mathbf{v}$, that is, the Markov chain forgets its past.

That is, the chain will convergence to the unique stationary distribution

Markov Chains

Definition

Irreducibility:

For each pairs of states *(i,j)*, there is a positive probability, starting in state *i*, that the process will ever enter state *j*.

= The matrix *T* cannot be reduced to separate smaller matrices

= Transition graph is connected.

It is possible to get to any state from any state.

Markov Chains

Definition

Aperiodicity: The chain cannot get trapped in cycles.

Definition

A state *i* has period *k* if any return to state *i,* must occur in multiples of *k* time steps. Formally, the period of a state *i* is defined as

$$k = \gcd\{n : \Pr(X_n = i | X_0 = i) > 0\}$$

(where "gcd" is the greatest common divisor)

For example, suppose it is possible to return to the state in {6,8,10,12,...} time steps. Then k=2

Markov Chains

Definition

Aperiodicity: The chain cannot get trapped in cycles.

In other words,

a state *i* is aperiodic if there exists n such that for all n' ≥ n,

$$\Pr(X_{n'} = i | X_0 = i) > 0$$

Definition

A Markov chain is aperiodic if every state is aperiodic.

Markov Chains

Example for periodic Markov chain:

Let $\mathrm{T} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$

In this case

$$(1/2, 1/2)T = (1/2, 1/2)$$

If we start the chain from (1,0), or (0,1), then the chain get traps into a cycle, it doesn't forget its past.

It has stationary distribution, but no limiting distribution!

Reversible Markov chains (Detailed Balance Property)

How can we find the limiting distribution of an irreducible and aperiodic Markov chain?

Definition: reversibility /detailed balance condition:

$$\pi_i P_{ij} = \pi_j P_{ji}, \ \forall (i, j)$$

Theorem:

A sufficient, but not necessary, condition to ensure that a particular π is the desired invariant distribution of the Markov chain is the detailed balance condition.

How fast can Markov chains forget the past?

MCMC samplers are

- irreducible and aperiodic Markov chains
- have the target distribution as the invariant distribution.
- the detailed balance condition is satisfied.

It is also important to design samplers that converge quickly.

The Hastings-Metropolis Algorithm

The Hastings-Metropolis Algorithm

Let $b_1, \ldots, b_m > 0$, and $B = \sum_{j=1}^{m} b_j$

Assume that m is so big, that it is difficult to calculate B.

Our goal:

Generate samples from the following **discrete** distribution:

$$P(X = j) = \pi_j = \frac{b_j}{B}$$

We don't know *B*!

The main idea is to construct a time-reversible Markov chain with $(\pi_1,\ldots,\pi_m)$ limit distributions

Later we will discuss what to do when the distribution is continuous

The Hastings-Metropolis Algorithm

Let {1,2,…,m} be the state space of a Markov chain that we can simulate.

Let $q(i,j) = p(j|i)$

Let $\{X_0, X_1, \ldots X_n, \ldots\}$ Markov chain be defined as follows:

$$\Pr(X_n = j \,|X_{n-1} = i) =$$

1., from state i go to state j with prob. $q(i,j)$

2., $\begin{cases} \text{with prob } 1-\alpha(i,j) \text{ go back to state } i, \\ \text{with prob } \alpha(i,j) \text{ stay in state } j. \end{cases}$

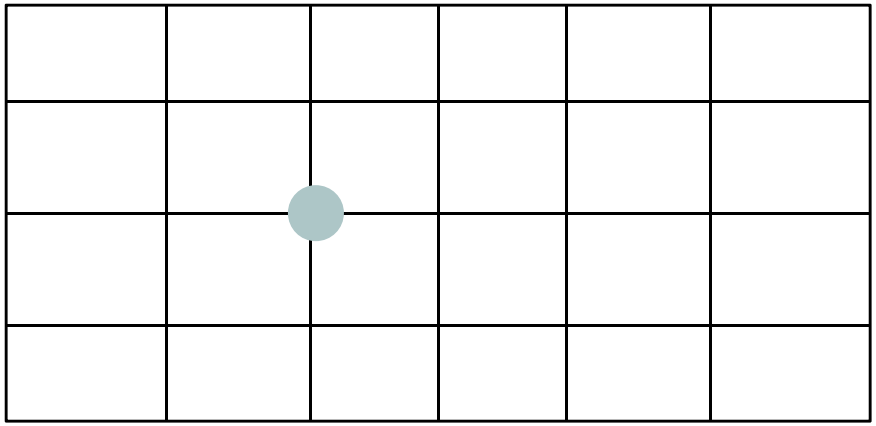
No rejection: we use all X_1, X_2,… X_n, …

Example for Large State Space

Let {1,2,…,m} be the state space of a Markov chain that we can simulate. Let $q(i,j) = p(j|i)$

d-dimensional grid:

- Max 2d possible movements at each grid point (linear in d)
- Exponentially large state space in dimension d



The Hastings-Metropolis Algorithm

$$\Pr(X_n = j \,|X_{n-1} = i) =$$

1., from state i go to state j with prob. $q(i,j)$

2., $\begin{cases} \text{with prob } 1-\alpha(i,j) \text{ go back to state } i, \\ \text{with prob } \alpha(i,j) \text{ stay in state } j. \end{cases}$

Theorem

$$\begin{aligned} P(X_{n+1} = j | X_n = i) &= q(i,j)\alpha(i,j) \quad \forall j \neq i \\ P(X_{n+1} = i | X_n = i) &= q(i,i) + \sum_{k \neq i} q(i,k)(1-\alpha(i,k)) \end{aligned}$$

Proof

We can go to state i from state i and also from other states $k \neq i$.

The Hastings-Metropolis Algorithm

Observation

$$\pi_i P_{ij} = \pi_j P_{ji} \quad \forall j \neq i \Leftrightarrow \pi_i q(i,j)\alpha(i,j) = \pi_j q(j,i)\alpha(j,i) \quad \forall j \neq i \quad (*)$$

Proof: $P_{ij} = P(X_{n+1} = j | X_n = i) = q(i,j)\alpha(i,j) \ \forall j \neq i$

Corollary

$$\Rightarrow \begin{cases} X_0, X_1, \ldots X_n, \ldots \text{ time reversible Markov chain} \\ \exists \pi_1, \ldots \pi_m \text{ stationary distribution} \end{cases}$$

Theorem

$$\text{If } \alpha(i,j) = \min\left(\frac{\pi_j q(j,i)}{\pi_i q(i,j)}, 1\right) = \min\left(\frac{b(j)q(j,i)}{b(i)q(i,j)}, 1\right) \Rightarrow$$

$$\Rightarrow \ (*) \text{ holds}$$

$$\Rightarrow \ (\pi_1, \ldots \pi_m) \text{ stationary distribution}$$

The Hastings-Metropolis Algorithm

$$\pi_i P_{ij} = \pi_j P_{ji} \quad \forall j \neq i \Leftrightarrow \pi_i q(i,j)\alpha(i,j) = \pi_j q(j,i)\alpha(j,i) \quad \forall j \neq i \quad (*)$$

Theorem

$$\text{If } \alpha(i,j) = \min\left(\frac{\pi_j q(j,i)}{\pi_i q(i,j)}, 1\right) = \min\left(\frac{b(j)q(j,i)}{b(i)q(i,j)}, 1\right) \Rightarrow$$

$$\Rightarrow \quad (*) \text{ holds}$$

$$\Rightarrow \quad (\pi_1, \ldots \pi_m) \text{ stationary distribution}$$

Proof:

$$\text{If } \alpha(i,j) = \frac{\pi_j q(j,i)}{\pi_i q(i,j)} \Leftrightarrow \alpha(j,i) = 1$$

Note: To calculate $\alpha(i,j)$ we didn't need to use $B = \sum_{j=1}^{m} b(j)$.

The Hastings-Metropolis Algorithm

1) Let Q be a Markov chain with $q(i,j) = P(j|i)$ state transition probabilites. Assume that we can sample from $q(i,j) = P(j|i)$.

2) Let $1 \leq k \leq m$ arbitrary, $n = 0$, and $X_0 = k$.

3) Sample X^* according to $P(X^* = j) = q(X_n, j)$, $j = 1, \ldots, m$ distribution.
(Go from X_n to state j using Markov chain Q)

4) Let $u \sim U_{[0,1]}$ (With prob $\alpha(i,j)$ stay in $X^* = j$)

5) If $u < \dfrac{b(X)q(X, X_n)}{b(X_n)q(X_n, X)} \Rightarrow X_{n+1} = X^*$
else $\Rightarrow X_{n+1} = X_n$ (Otherwise go back)

6) $n = n + 1$

7) Back to 3

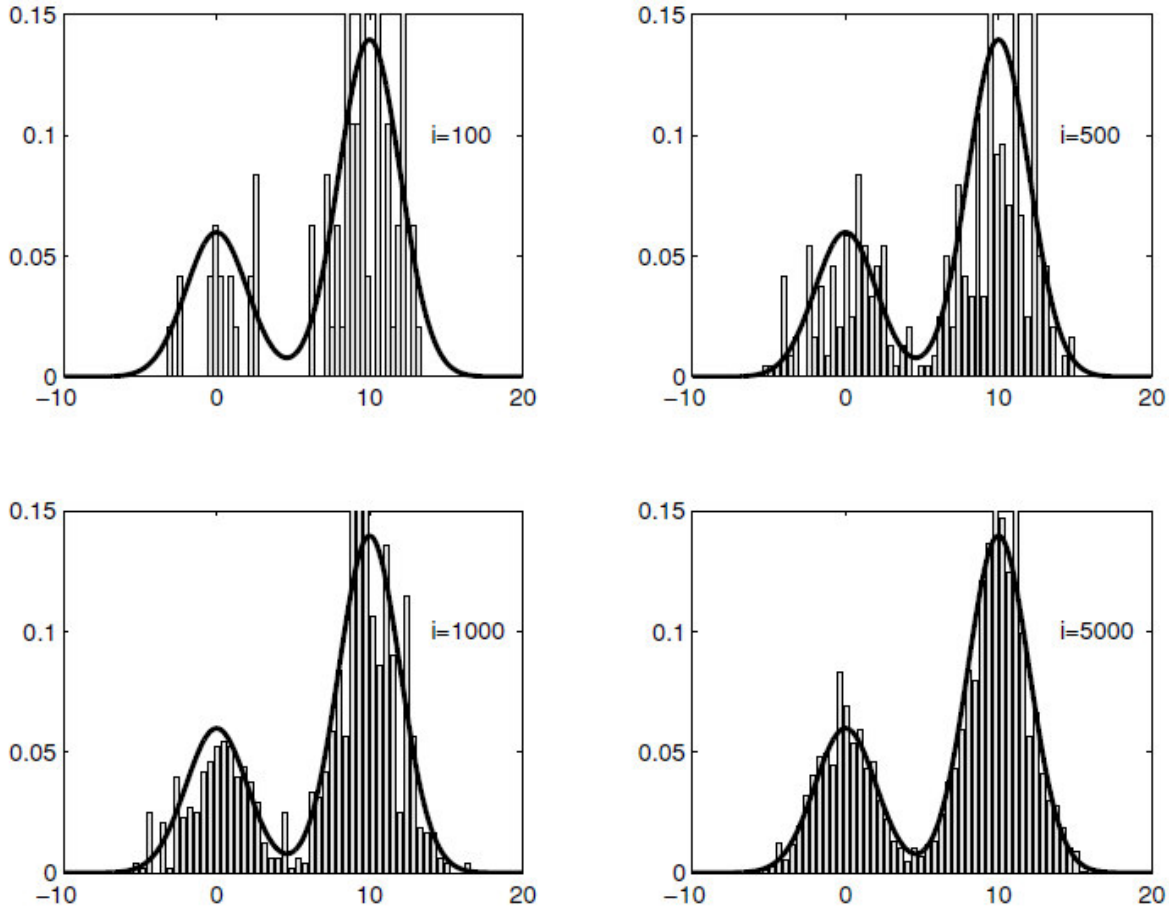
It is not rejection sampling, we use all the samples!

Continuous Distributions

- The same algorithm can be used for continuous distributions as well.
- In this case, the state space is continuous.

Experiment with HM

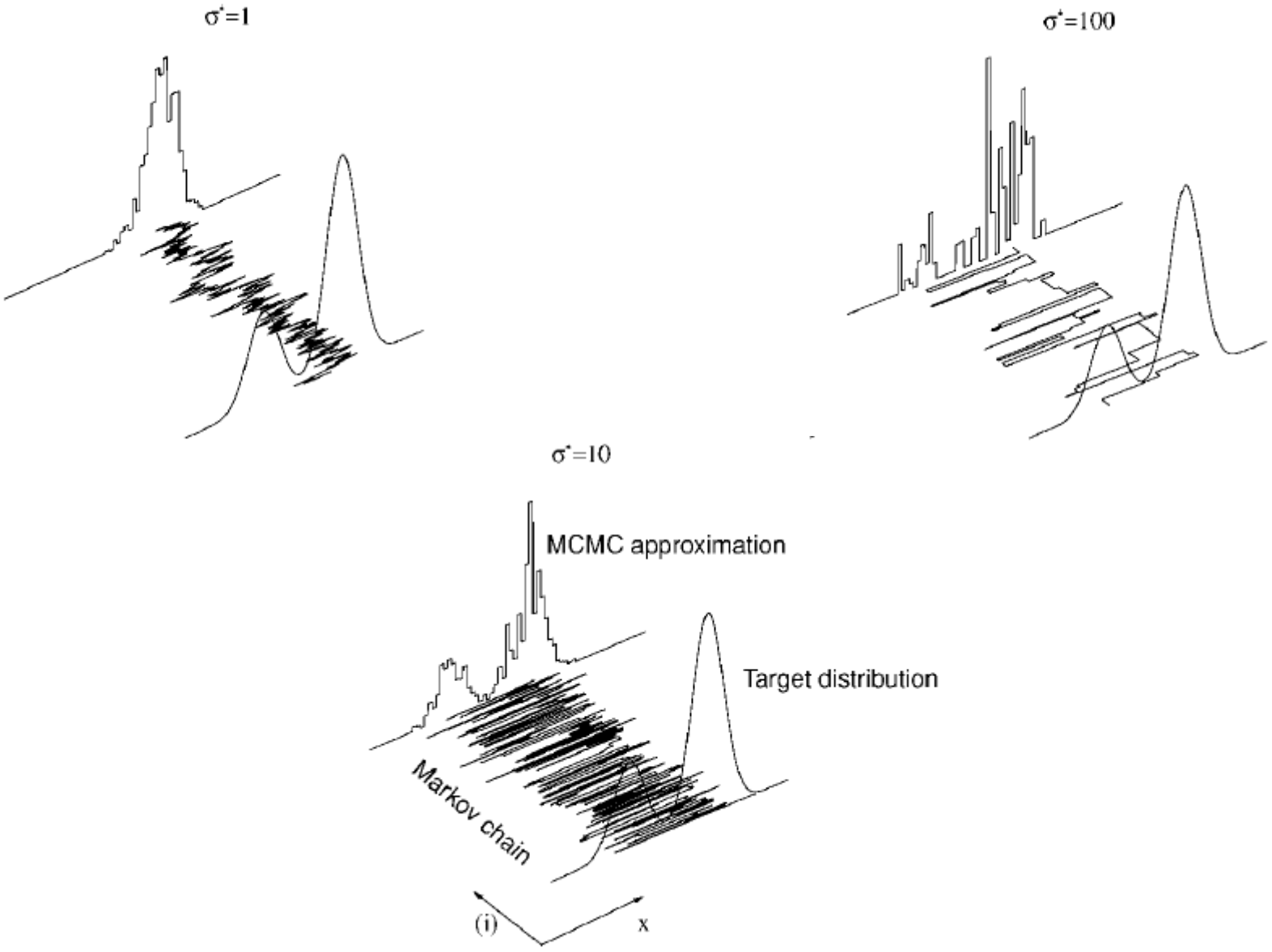
An application for continuous distributions



Bimodal target distribution: $p(x) \propto 0.3 \exp(-0.2x^2) + 0.7 \exp(-0.2(x-10)^2)$

$q(x \mid x^{(i)}) = N(x^{(i)}, 100)$, 5000 iterations

Good proposal distrib. is important



HM on Combinatorial Sets

Let $\mathcal{P} = \{x_1, \ldots x_n \,|\, x_1, \ldots x_n$ is a permuation of $(1, \ldots, n)$ such that, $\sum_{j=1}^{n} j x_j > a\}$, where a is a given constant.

Generate **uniformly distributed** samples from the set of permutations $\mathcal{P}$.

Let n=3, and a=12:

{1,2,3}: 1+4+9=14

{1,3,2}: 1+6+6=13

{2,3,1}: 2+6+3=11

{2,1,3}: 2+2+9=13

{3,1,2}: 3+2+6=11

{3,2,1}: 3+4+3=10

HM on Combinatorial Sets

To define a simple Markov chain on $\mathcal{P}$, we need the concept of **neighboring elements** (permutations):

Definition: Two permutations are **neighbors**, if one results from the interchange of two of the positions of the other:

(1,2,3,4) and (1,2,4,3) are neighbors.

(1,2,3,4) and (1,3,4,2) are not neighbors.

HM on Combinatorial Sets

Let $N(i)$ be the number set of state i.

$$\text{Let } q(i,j) = P(j|i) = \begin{cases} \frac{1}{|N(i)|} & \text{if } j \in N(i). \\ 0 & \text{Otherwise} \end{cases}$$

$$\alpha(i,j) = \min\left(\frac{\pi_j q(j,i)}{\pi_i q(i,j)}, 1\right) = \min\left(\frac{1\frac{1}{N(j)}}{1\frac{1}{N(i)}}, 1\right) = \min\left(\frac{N(i)}{N(j)}, 1\right)$$

$\Rightarrow$ the limit distribution of the Markov chain is uniform over $\mathcal{P}$ with probabilities $\frac{1}{|\mathcal{P}|}$

That is what we wanted!

Gibbs Sampling: The Problem

Let $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$

Let $p(x_1, \ldots, x_n) \geq 0$ be a non-normalized distribution ($\int p(x) \neq 1$, $p(x) \geq 0$), and let A be a complicated set.

Suppose that we can generate samples from

$$P(X_i = x | X_j = x_j, \forall j \neq i)$$

e.g. $P(X_3 = x_3 | X_1 = x_1, X_2 = x_2, X_4 = x_4, X_5 = x_5)$

Our goal is to generate samples from

$$f(x_1, \ldots, x_n) = \begin{cases} 0 & \text{if } \mathbf{x} \notin A \\ \frac{p(\mathbf{x})}{p(\mathbf{x} \in A)} & \text{if } \mathbf{x} \in A \end{cases}$$

Gibbs Sampling: Pseudo Code

1. We are in $\mathbf{x} = (x_1, \ldots, x_n) \in A$

2. Draw a random state $i \in \{1, \ldots, n\}$ with prob. $1/n$.

3. Sample x from $x \sim P(X_i = x | X_j = x_j, \forall j \neq i)$.

4. Let $\mathbf{y} = (x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n)$

5. If

   $(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n) \in \mathbf{A} \Rightarrow x_i = x$, accept this new state

   $(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n) \notin \mathbf{A} \Rightarrow x_i$ stays in the old x_i

6. New sample point: $(x_1, \ldots, x_n)$. Go back to 2

Gibbs Sampling: Theory

Consider the following HM sampler:

Let

$$\begin{aligned} q(\mathbf{x},\mathbf{y}) &= q(\overbrace{(x_1,\ldots,x_n)}^{\mathbf{x}},\overbrace{(x_1,\ldots,x_{i-1},x,x_{i+1},x_n)}^{\mathbf{y}}) \\ &\doteq \frac{1}{n}P(X_i = x | X_j = x_j, \forall j \neq i) \\ &= \frac{1}{n}\frac{P(\mathbf{y})}{P(X_j = x_j, \forall j \neq i)} \end{aligned}$$

and let $$\alpha(\mathbf{x},\mathbf{y}) = \min\left(\frac{f(\mathbf{y})q(\mathbf{y},\mathbf{x})}{f(\mathbf{x})q(\mathbf{x},\mathbf{y})}, 1\right)$$

Observation: By construction, this **HM sampler** would sample from

$$f(x_1,\ldots,x_n) = \begin{cases} 0 & \text{if } \mathbf{x} \notin A \\ \frac{p(\mathbf{x})}{p(\mathbf{x}\in A)} & \text{if } \mathbf{x} \in A \end{cases}$$

We will prove that this HM sampler = Gibbs sampler.

Gibbs Sampling is a Special HM

Theorem: The Gibbs sampling is a special case of HM with

$$\alpha(\mathbf{x},\mathbf{y}) = \begin{cases} 1 & \text{if } \mathbf{x} \in A, \mathbf{y} \in A \\ 0 & \text{if } \mathbf{x} \in A, \mathbf{y} \notin A \end{cases}$$

Proof:

By definition: $f(x_1,\ldots,x_n) = \begin{cases} 0 & \text{if } \mathbf{x} \notin A \\ \frac{p(\mathbf{x})}{p(\mathbf{x}\in A)} & \text{if } \mathbf{x} \in A \end{cases}$

If $\mathbf{x} \in A, \mathbf{y} \in A \Rightarrow \dfrac{f(\mathbf{y})q(\mathbf{y},\mathbf{x})}{f(\mathbf{x})q(\mathbf{x},\mathbf{y})} = \dfrac{\frac{p(\mathbf{y})}{p(\mathbf{y}\in A)}q(\mathbf{y},\mathbf{x})}{\frac{p(\mathbf{x})}{p(\mathbf{x}\in A)}q(\mathbf{x},\mathbf{y})} =$

$$= \frac{p(\mathbf{y})q(\mathbf{y},\mathbf{x})}{p(\mathbf{x})q(\mathbf{x},\mathbf{y})} = \frac{p(\mathbf{y})\frac{1}{n}\frac{P(\mathbf{x})}{P(Y_j=y_j, j\neq i)}}{p(\mathbf{x})\frac{1}{n}\frac{P(\mathbf{y})}{P(X_j=x_j, j\neq i)}} = \frac{p(\mathbf{y})p(\mathbf{x})}{p(\mathbf{x})p(\mathbf{y})} = 1$$

since $P(X_j = x_j, j \neq i) = P(Y_j = y_j, j \neq i)$

Gibbs Sampling is a Special HM

Proof:

$$\text{If} \quad \mathbf{x} \in A, \mathbf{y} \notin A \Rightarrow \frac{f(\mathbf{y})q(\mathbf{y},\mathbf{x})}{f(\mathbf{x})q(\mathbf{x},\mathbf{y})} = \frac{0q(\mathbf{y},\mathbf{x})}{\frac{p(\mathbf{x})}{p(\mathbf{x}\in A)}q(\mathbf{x},\mathbf{y})} = 0$$

$$\text{since } \alpha(\mathbf{x},\mathbf{y}) = \begin{cases} 1 & \text{if } \mathbf{x} \in A, \mathbf{y} \in A \\ 0 & \text{if } \mathbf{x} \in A, \mathbf{y} \notin A \end{cases}$$

Gibbs Sampling in Practice