Chapter 11: Sampling Methods

Lei Tang

Department of CSE
Arizona State University

Dec. 18th, 2007

Outline

MCMC

- We've discussed the rejection sampling and importance sampling to find expectations of a function.
- They suffer from severe limitations particularly in spaces of high dimensionality.
- We now discuss a very general and powerful framework called *Markov Chain Monte Carlo* (MCMC).
- MCMC methods have their origin in physics and started to have a significant impact on the field of statistics at the end of 1980s.

Basic setup

- Similar to rejection and importance sampling, we again sample from a proposal distribution.
- We maintain current state $\mathbf{z}^{(\tau)}$, and the proposal distribution $q(\mathbf{z}|\mathbf{z}^{(\tau)})$ depends on current state.
- So the sequence $\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \cdots$ forms a Markov chain (the next sample depends on the previous one).
- Assumption: $p(\mathbf{z}) = \hat{p}(\mathbf{z})/Z_p$ where Z_p is unknown and $\hat{p}(\mathbf{z})$ is easy to compute.
- The proposal distribution should be straightforward to draw samples.
- Each cycle we generate a sample $\mathbf{z}^*$ and accept it with proper criteria.

Metropolis Algorithm

- Assume the proposal distribution is symmetric:

$$q(\mathbf{z}_A|\mathbf{z}_B) = q(\mathbf{z}_B|\mathbf{z}_A)$$

- The candidate sample z^* is accepted with probability

$$A(\mathbf{z}^*, \mathbf{z}^{(\tau)}) = min\left(1, \frac{\hat{p}(\mathbf{z}*)}{\hat{p}(\mathbf{z}^{(\tau)})}\right)$$

  This can be done by choosing a random number u from a uniform distribution over $(0, 1)$, and accepting the sample if $A(\mathbf{z}^*, \mathbf{z}^{(\tau)}) > u$.

- $$\mathbf{z}^{(\tau+1)} = \begin{cases} \mathbf{z}^* & \text{if accepted} \\ \mathbf{z}^{(\tau)} & \text{if rejected} \end{cases}$$

  If p(z*) is large, it's likely to be accepted.

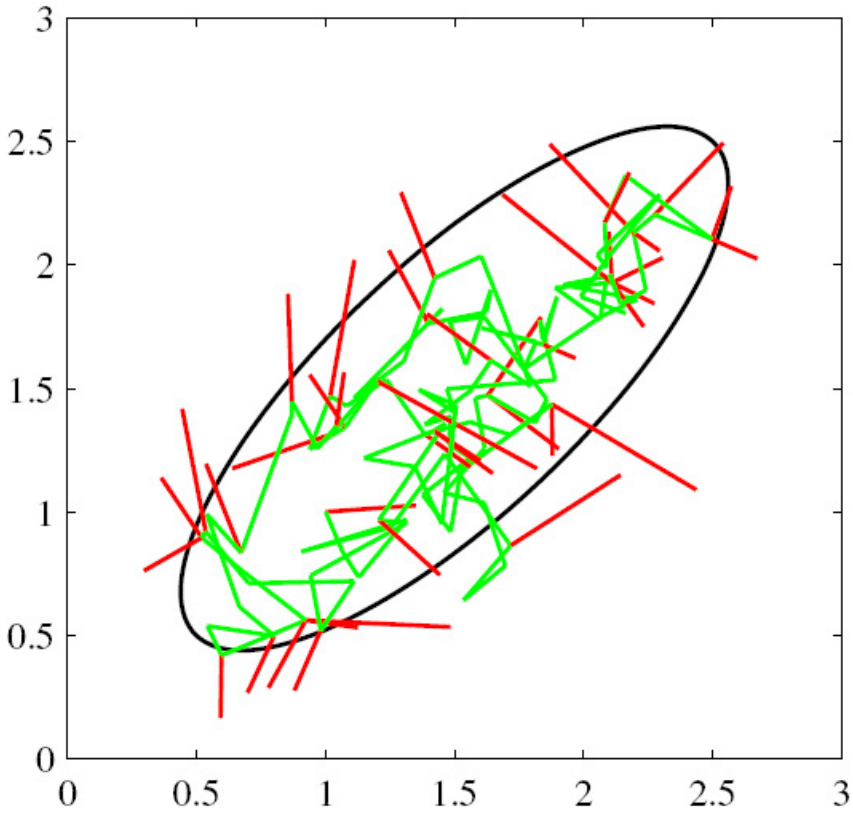
- As long as $q(\mathbf{z}_A|\mathbf{z}_B) > 0$, the distribution of $\mathbf{z}^{(\tau)} \rightarrow p(z)$ when $\tau \rightarrow \infty$. (We'll prove this later)

How to handle dependence?

- The sequence $\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \cdots$, is not independent.
- Usually, discard the most of the sequence and retain every M^{th} sample.
- Sometimes, need to throw away the first few hundreds samples if you start from a not-so-good initial point (to avoid the burn-in period)

An Example

A simple illustration using Metropolis algorithm to sample from a Gaussian distribution whose one standard-deviation contour is shown by the ellipse. The proposal distribution is an isotropic Gaussian distribution whose standard deviation is 0.2. Steps that are accepted are shown as green lines, and rejected steps are shown in red. A total of 150 candidate samples are generated, of which 43 are rejected.



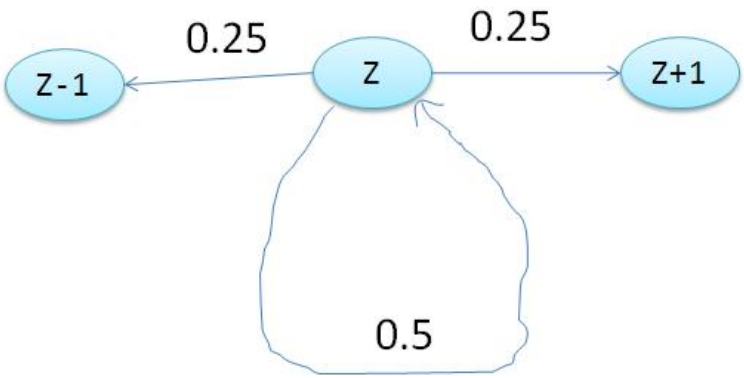
- The proposal distribution is Gaussian whose standard deviation is 0.2. Clearly, $q(\mathbf{z}_A|\mathbf{z}_B) = q(\mathbf{z}_B|\mathbf{z}_A)$.
- Each step search in the space of a rectangle, but favor the samples toward high-density.

Common Questions

1. Why does Metropolis Algorithm work?
2. How efficient?
3. Is it possible to relax the symmetry property of proposal distribution?

Random Walk: Blind?

To investigate the property of MCMC, we look at a specific example of random walk first:



$$
\begin{aligned}
p(z^{(\tau+1)} = z^{(\tau)}) &= 0.5 \\
p(z^{(\tau+1)} = z^{(\tau)} + 1) &= 0.25 \\
p(z^{(\tau+1)} = z^{(\tau)} - 1) &= 0.25
\end{aligned}
$$

If start from $z^{(0)} = 0$, then $E[z^{(\tau)}] = 0$;

Quiz: how to prove this?

$$
\begin{aligned}
E[z^{(\tau+1)}] &= 0.5E[z^{(\tau)}] + 0.25(E[z^{(\tau)}] + 1) + 0.25(E[z^{(\tau)}] - 1) \\
&= E(z^{(\tau)})
\end{aligned}
$$

Random Walk is Inefficient

How to measure the average distance between starting and ending points?

$$E[(z^{(\tau)})^2] = \frac{\tau}{2}$$

$$\begin{aligned} E[(z^{(\tau+1)})^2] &= 0.5E[(z^{(\tau)})^2] + \\ &\quad 0.25(E[(z^{(\tau)})^2] + 2E[z^{(\tau)}] + 1) + \\ &\quad 0.25(E[(z^{(\tau)})^2] - 2E[z^{(\tau)}] + 1) \\ &= E[(z^{(\tau)})^2] + 0.5 \\ \Longrightarrow E[(z^{(\tau)})^2] &= \frac{\tau}{2} \end{aligned}$$

- The average distance between start and ending points of τ steps is $O(\sqrt{\tau})$.
- Random walk is very inefficient in exploring the state space.
- A central goal of MCMC is to avoid random walk behavior.

Random Walk is Inefficient

How to measure the average distance between starting and ending points?

$$E[(z^{(\tau)})^2] = \frac{\tau}{2}$$

$$\begin{aligned} E[(z^{(\tau+1)})^2] &= 0.5E[(z^{(\tau)})^2] + \\ &\quad 0.25(E[(z^{(\tau)})^2] + 2E[z^{(\tau)}] + 1) + \\ &\quad 0.25(E[(z^{(\tau)})^2] - 2E[z^{(\tau)}] + 1) \\ &= E[(z^{(\tau)})^2] + 0.5 \\ \Longrightarrow E[(z^{(\tau)})^2] &= \frac{\tau}{2} \end{aligned}$$

- The average distance between start and ending points of τ steps is $O(\sqrt{\tau})$.
- Random walk is very inefficient in exploring the state space.
- A central goal of MCMC is to avoid random walk behavior.

Random Walk is Inefficient

How to measure the average distance between starting and ending points?

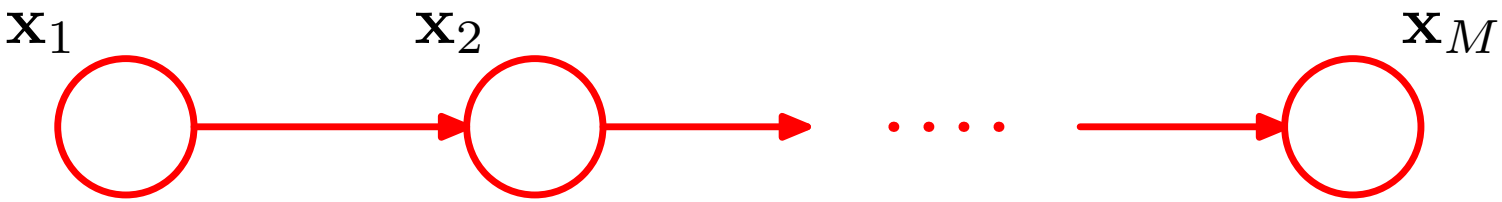
$$E[(z^{(\tau)})^2] = \frac{\tau}{2}$$

$$\begin{aligned} E[(z^{(\tau+1)})^2] &= 0.5E[(z^{(\tau)})^2] + \\ &\quad 0.25(E[(z^{(\tau)})^2] + 2E[z^{(\tau)}] + 1) + \\ &\quad 0.25(E[(z^{(\tau)})^2] - 2E[z^{(\tau)}] + 1) \\ &= E[(z^{(\tau)})^2] + 0.5 \\ \Longrightarrow E[(z^{(\tau)})^2] &= \frac{\tau}{2} \end{aligned}$$

- The average distance between start and ending points of τ steps is $O(\sqrt{\tau})$.
- Random walk is very inefficient in exploring the state space.
- A central goal of MCMC is to avoid random walk behavior.

Markov Chain

$$p(\mathbf{z}^{(m+1)}|\mathbf{z}^{(1)}, \cdots, \mathbf{z}^{(m)}) = p(\mathbf{z}^{(m+1)}|\mathbf{z}^{(m)})$$



- *Transition Probabilities*: $T_m(\mathbf{z}^{(m)}, \mathbf{z}^{(m+1)}) = p(\mathbf{z}^{(m+1)}|\mathbf{z}^{(m)})$.
- A Markov chain is independent is *homogeneous* if the transition probability are the same for $\forall m$.
- The marginal distribution:

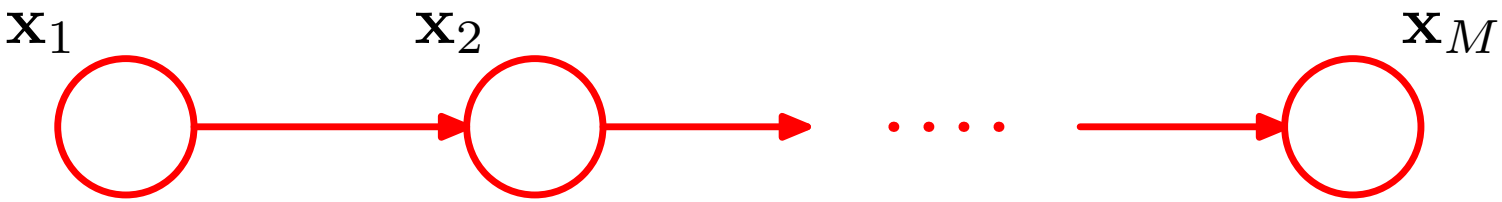
$$p(\mathbf{z}^{(m+1)}) = \sum_{\mathbf{z}^{(m)}} p(\mathbf{z}^{(m+1)}|\mathbf{z}^{(m)}) p(\mathbf{z}^{(m)}))$$

- Stationary(invariant) distribution: each step in the chain leaves the distribution invariant.

$$p^*(\mathbf{z}) = \sum_{\mathbf{z}'} T(\mathbf{z}', \mathbf{z}) p^*(\mathbf{z}')$$

Markov Chain

$$p(\mathbf{z}^{(m+1)}|\mathbf{z}^{(1)}, \cdots, \mathbf{z}^{(m)}) = p(\mathbf{z}^{(m+1)}|\mathbf{z}^{(m)})$$



- *Transition Probabilities*: $T_m(\mathbf{z}^{(m)}, \mathbf{z}^{(m+1)}) = p(\mathbf{z}^{(m+1)}|\mathbf{z}^{(m)})$.
- A Markov chain is independent is *homogeneous* if the transition probability are the same for $\forall m$.
- The marginal distribution:

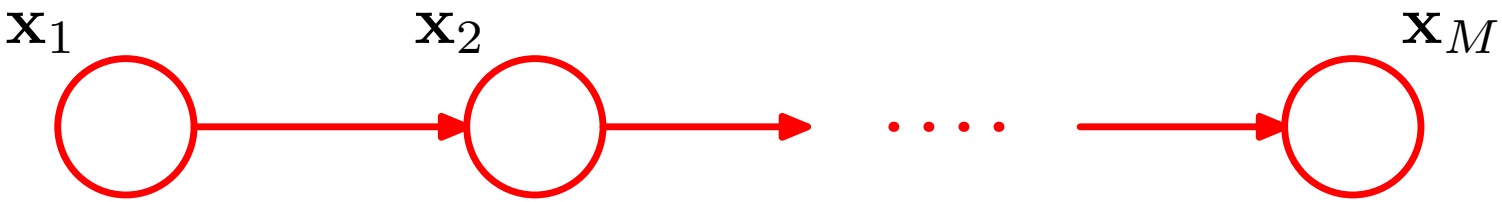
$$p(\mathbf{z}^{(m+1)}) = \sum_{\mathbf{z}^{(m)}} p(\mathbf{z}^{(m+1)}|\mathbf{z}^{(m)})p(\mathbf{z}^{(m)}))$$

- Stationary(invariant) distribution: each step in the chain leaves the distribution invariant.

$$p^*(\mathbf{z}) = \sum_{\mathbf{z}'} T(\mathbf{z}', \mathbf{z})p^*(\mathbf{z}')$$

Markov Chain

$$p(\mathbf{z}^{(m+1)}|\mathbf{z}^{(1)}, \cdots, \mathbf{z}^{(m)}) = p(\mathbf{z}^{(m+1)}|\mathbf{z}^{(m)})$$



- *Transition Probabilities*: $T_m(\mathbf{z}^{(m)}, \mathbf{z}^{(m+1)}) = p(\mathbf{z}^{(m+1)}|\mathbf{z}^{(m)})$.
- A Markov chain is independent is *homogeneous* if the transition probability are the same for $\forall m$.
- The marginal distribution:

$$p(\mathbf{z}^{(m+1)}) = \sum_{\mathbf{z}^{(m)}} p(\mathbf{z}^{(m+1)}|\mathbf{z}^{(m)})p(\mathbf{z}^{(m)}))$$

- Stationary(invariant) distribution: each step in the chain leaves the distribution invariant.

$$p^*(\mathbf{z}) = \sum_{\mathbf{z}'} T(\mathbf{z}', \mathbf{z})p^*(\mathbf{z}')$$

Detailed Balance

- A *sufficient* (but not necessary) condition for ensuring the required distribution to be invariant is

$$p^*(\mathbf{z})T(\mathbf{z},\mathbf{z}') = p^*(\mathbf{z}')T(\mathbf{z}',\mathbf{z})$$

  This property is called detailed balance.

- A Markov chain satisfy the detailed balance will leave the distribution invariant:

$$\begin{aligned}
&\sum_{\mathbf{z}'} p^*(\mathbf{z}')T(\mathbf{z}',\mathbf{z}) \\
=\ &\sum_{\mathbf{z}'} p^*(\mathbf{z})T(\mathbf{z},\mathbf{z}') \quad \text{(Property of detailed balance)} \\
=\ &p^*(\mathbf{z})\sum_{\mathbf{z}'} p(\mathbf{z}'|\mathbf{z}) \\
=\ &p^*(\mathbf{z}) \quad (\sum_{\mathbf{z}'} p(\mathbf{z}'|\mathbf{z}) = 1)
\end{aligned}$$

Detailed Balance

- A *sufficient* (but not necessary) condition for ensuring the required distribution to be invariant is

$$p^*(\mathbf{z})\,T(\mathbf{z},\mathbf{z}') = p^*(\mathbf{z}')\,T(\mathbf{z}',\mathbf{z})$$

  This property is called detailed balance.
- A Markov chain satisfy the detailed balance will leave the distribution invariant:

$$\begin{aligned}
& \sum_{\mathbf{z}'} p^*(\mathbf{z}')\,T(\mathbf{z}',\mathbf{z}) \\
= & \sum_{\mathbf{z}'} p^*(\mathbf{z})\,T(\mathbf{z},\mathbf{z}') \quad \text{(Property of detailed balance)} \\
= & \; p^*(\mathbf{z}) \sum_{\mathbf{z}'} p(\mathbf{z}'|\mathbf{z}) \\
= & \; p^*(\mathbf{z}) \quad \Big(\sum_{\mathbf{z}'} p(\mathbf{z}'|\mathbf{z}) = 1\Big)
\end{aligned}$$

- A Markov chain satisfy the detailed balance is reversible.
- Detailed balance is *stronger* than the requirement of stationary distribution.
- Quiz: Can you give me a counter example?
- Our goal is to set up a Markov chain such that the invariant distribution is our desired distribution.

Ergodicity

- Goal: set up a Markov chain such that the invariant distribution is our desired distribution.
- We must require the ergodicity property: for $m \rightarrow \infty$, the distribution $p(\mathbf{z}^{(m)})$ converges to the required invariant distribution $p^*(z)$, irrespective of the initial choice. The invariant distribution is called the equilibrium distribution.
- Each ergodic Markov chain can have only one equilibrium distribution.
- It can be shown that a homogeneous Markov chain will be ergodic, subject only to weak restrictions on the invariant distribution and the transition probabilities.

Ergodicity

- Goal: set up a Markov chain such that the invariant distribution is our desired distribution.
- We must require the ergodicity property: for $m \rightarrow \infty$, the distribution $p(\mathbf{z}^{(m)})$ converges to the required invariant distribution $p^*(z)$, irrespective of the initial choice. The invariant distribution is called the equilibrium distribution.
- Each ergodic Markov chain can have only one equilibrium distribution.
- It can be shown that a homogeneous Markov chain will be ergodic, subject only to weak restrictions on the invariant distribution and the transition probabilities.

Ergodicity

- Goal: set up a Markov chain such that the invariant distribution is our desired distribution.
- We must require the ergodicity property: for $m \rightarrow \infty$, the distribution $p(\mathbf{z}^{(m)})$ converges to the required invariant distribution $p^*(z)$, irrespective of the initial choice. The invariant distribution is called the equilibrium distribution.
- Each ergodic Markov chain can have only one equilibrium distribution.
- It can be shown that a homogeneous Markov chain will be ergodic, subject only to weak restrictions on the invariant distribution and the transition probabilities.

Weak restriction for ergodicity

If a homogeneous Markov chain on a finite state space with transition probabilities $T(z, z')$ has π as an invariant distribution and

$$\nu = \min_{z} \min_{z':\pi(z')>0} \frac{T(z, z')}{\pi(z')} > 0$$

then the Markov chain is ergodic. i.e. regardless of initial probabilities, $p_0(z)$:

$$\lim_{n\to\infty} p_n(z) = \pi(z)$$

A bound on the rate of convergence is given by

$$|\pi(z) - p_n(z)| \leq (1 - \nu)^n$$

Proof Outline

- Decompose the distribution at time n as a "mixture" of the invariant distribution and another arbitrary distribution;
- The proportion with the invariant distribution approaches to 1 as n approaches to infinity.
- Specifically, $p_n(z) = [1 - (1 - \nu)^n]\pi(z) + (1 - \nu)^n r_n(z)$.
- The theorem can be proved by induction.

- $p_n(z) = [1 - (1 - \nu)^n]\pi(z) + (1 - \nu)^n r_n(z)$ is automatically satisfied when $n = 0$ (Just set $r_0(z) = p_0(z)$).

Proof

$$
\begin{aligned}
& p_{n+1}(z) \\
= & \sum_{z'} p_n(z')\, T(z', z) \\
= & [1-(1-\nu)^n] \sum_{z'} \pi(z')\, T(z', z) + (1-\nu)^n \sum_{z'} r_n(z')\, T(z', z) \\
= & [1-(1-\nu)^n]\pi(z) + (1-\nu)^n \sum_{z'} r_n(z')[T(z', z) - \nu\pi(z) + \nu\pi(z)] \\
= & [1-(1-\nu)^n]\pi(z) + (1-\nu)^n \nu\pi(z) \\
& \quad (1-\nu)^n \sum_{z'} r_n(z')[T(z', z) - \nu pi(z)] \\
= & [1-(1-\nu)^{n+1}]\pi(z) + (1-\nu)^{n+1} \sum_{z'} r_n(z') \frac{T(z', z) - \nu\pi(z)}{1-\nu} \\
= & [1-(1-\nu)^{n+1}]\pi(z) + (1-\nu)^{n+1} r_{n+1}(z)
\end{aligned}
$$

$$r_{n+1}(z) = \sum_{z'} r_n(z') \frac{T(z', z) - \nu\pi(z)}{1 - \nu} \quad \text{must be a distribution}$$

$$\begin{aligned} \sum_{z} r_{n+1}(z) &= \sum_{z,z'} r_n(z') \frac{T(z', z) - \nu\pi(z)}{1 - \nu} \\ &= \frac{\sum_{z,z'} r_n(z') T(z', z) - \sum_{z} \nu\pi(z)}{1 - \nu} = 1 \end{aligned}$$

Note that $r_n(z) \geq 0$ as long as

$$\nu = \min_{z'} \min_{z:\pi(z)>0} \frac{T(z', z)}{\pi(z)} > 0$$

$$T(z', z) - \nu\pi(z) \geq 0$$

Hence, $r_n(z)$ is a valid distribution.
Note that $\nu \leq 1$.

- A special case to satisfy the theorem is that

$$T(z', z) \geq 0 \text{ for } \forall z, z'$$

- The above theorem applies only to homogeneous Markov chains.
- But the algorithm we'll discuss is not homogeneous, but of a simple cyclic type, in which $T_n = T_{n+d}$.
- We can look at the state only at times that are multiple of d, then we see a homogeneous Markov chain, with transition matrix $T_0 T_1 \cdots T_{d-1}$.
- If for a homogeneous Markov chain, condition does not hold for one step but hold for the k-step transition probabilities T^k. It's also sufficient.
- How about contiguous version?

- A special case to satisfy the theorem is that

$$T(z', z) \geq 0 \text{ for } \forall z, z'$$

- The above theorem applies only to homogeneous Markov chains.
- But the algorithm we'll discuss is not homogeneous, but of a simple cyclic type, in which $T_n = T_{n+d}$.
- We can look at the state only at times that are multiple of d, then we see a homogeneous Markov chain, with transition matrix $T_0 T_1 \cdots T_{d-1}$.
- If for a homogeneous Markov chain, condition does not hold for one step but hold for the k-step transition probabilities T^k. It's also sufficient.
- How about contiguous version?

- A special case to satisfy the theorem is that

$$T(z', z) \geq 0 \text{ for } \forall z, z'$$

- The above theorem applies only to homogeneous Markov chains.
- But the algorithm we'll discuss is not homogeneous, but of a simple cyclic type, in which $T_n = T_{n+d}$.
- We can look at the state only at times that are multiple of d, then we see a homogeneous Markov chain, with transition matrix $T_0 T_1 \cdots T_{d-1}$.
- If for a homogeneous Markov chain, condition does not hold for one step but hold for the k-step transition probabilities T^k. It's also sufficient.
- How about contiguous version?

- A special case to satisfy the theorem is that

$$T(z', z) \geq 0 \text{ for } \forall z, z'$$

- The above theorem applies only to homogeneous Markov chains.
- But the algorithm we'll discuss is not homogeneous, but of a simple cyclic type, in which $T_n = T_{n+d}$.
- We can look at the state only at times that are multiple of d, then we see a homogeneous Markov chain, with transition matrix $T_0 T_1 \cdots T_{d-1}$.
- If for a homogeneous Markov chain, condition does not hold for one step but hold for the k-step transition probabilities T^k. It's also sufficient.
- How about contiguous version?

Construction of Transition Probabilities

In practice, we often construct transition probabilities from a set of 'base' transitions $B_1, B_2, \cdots, B_K$.

$$T(\mathbf{z}', \mathbf{z}) = \sum_{k=1}^{K} \alpha_k B_k(\mathbf{z}', \mathbf{z}) \quad (1)$$

$$T(\mathbf{z}', \mathbf{z}) = \sum_{z_1} \cdots \sum_{z_{K-1}} B_1(\mathbf{z}', z_1) \cdots B_{K-1}(z_{K-2}, z_{K-1}) B_K(z_{K-1}, \mathbf{z}) \quad (2)$$

- If the distribution is invariant with respect to base transitions, it will be invariant for both (1) and (2).
- If base transitions satisfy detailed balance, then T in (1) also satisfy satisfy detailed balance. But T in (2) does not hold.
- A common example of the use of composite transition probabilities is where base transition changes only a subset of variables.

Metropolis-Hastings algorithm

- In previous Metropolis algorithm, the proposal distribution must be symmetric.
- Here is a generalization where the proposal distribution can be asymmetric.
- Only need to change the acceptance criterion to

$$A_k(\mathbf{z}^*, \mathbf{z}^{(\tau)}) = \min\left(1, \frac{\hat{p}(\mathbf{z}^*) q_k(\mathbf{z}^{(\tau)}|\mathbf{z}^*)}{\hat{p}(\mathbf{z}^{(\tau)}) q_k(\mathbf{z}^*|\mathbf{z}^{(\tau)})}\right)$$

- This evaluation does not require knowledge of normalization.
- For a symmetric proposal distribution, the Metropolis-Hastings algorithm reduces to Metropolis algorithm.

Metropolis-Hastings algorithm

- In previous Metropolis algorithm, the proposal distribution must be symmetric.
- Here is a generalization where the proposal distribution can be asymmetric.
- Only need to change the acceptance criterion to

$$A_k(\mathbf{z}^*, \mathbf{z}^{(\tau)}) = \min\left(1, \frac{\hat{p}(\mathbf{z}^*) q_k(\mathbf{z}^{(\tau)}|\mathbf{z}^*)}{\hat{p}(\mathbf{z}^{(\tau)}) q_k(\mathbf{z}^*|\mathbf{z}^{(\tau)})}\right)$$

- This evaluation does not require knowledge of normalization.
- For a symmetric proposal distribution, the Metropolis-Hastings algorithm reduces to Metropolis algorithm.

Invariant distribution

$$A_k(\mathbf{z}', \mathbf{z}) = \min\left(1, \frac{\hat{p}(\mathbf{z}')q_k(\mathbf{z}|\mathbf{z}')}{\hat{p}(\mathbf{z})q_k(\mathbf{z}'|\mathbf{z})}\right)$$

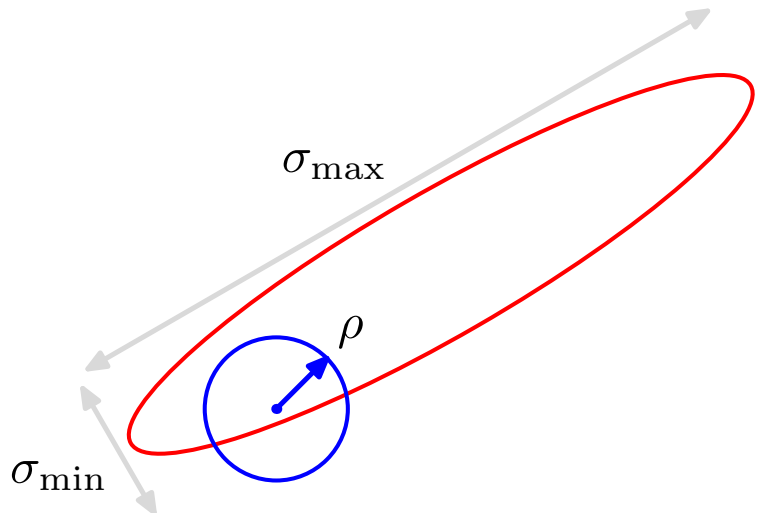
We show that $p(\mathbf{z})$ is an invariant distribution of the Markov chain, by showing the detailed balance is satisfied.

$$\begin{aligned}
p(z)T(z, z') &= p(z)q_k(z'|z)A_k(z', z) \\
&= \min(p(z)q_k(z'|z), p(z')q_k(z|z')) \\
&= \min(p(z')q_k(z|z'), p(z)q_k(z'|z)) \\
&= p(z')q_k(z|z')A_k(z, z') \\
&= p(z')T(z', z)
\end{aligned}$$

How about those samples being kept?

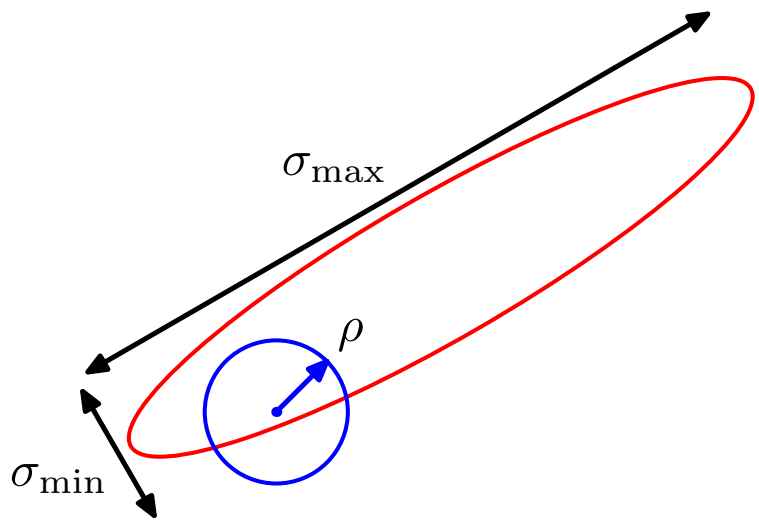
Proposal distribution

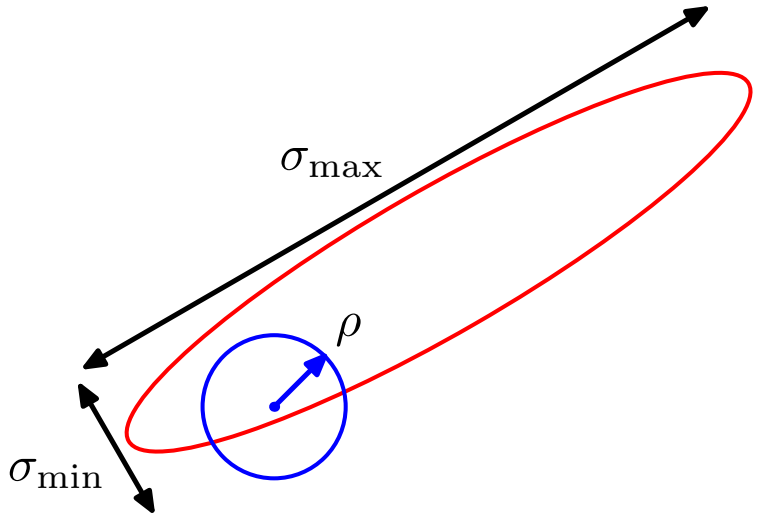
- The proposal distribution can affect the performance of sampling.
- For continuous state space, a common choice is a Gaussian centered on the current state.
- If the variance is small, the proportion of accepted transitions will be high, but the progress through the state space takes the form a slow random walk.
- If the parameter is large, then the rejection rate will be high, many of the proposed steps will be to states for which the probability $p(z)$ is low.



Proposal distribution

- The proposal distribution can affect the performance of sampling.
- For continuous state space, a common choice is a Gaussian centered on the current state.
- If the variance is small, the proportion of accepted transitions will be high, but the progress through the state space takes the form a slow random walk.
- If the parameter is large, then the rejection rate will be high, many of the proposed steps will be to states for which the probability $p(z)$ is low.





- This suggest that ρ should be of the same order as the smallest length scale σ_{min}. However, for those extended dimensions, the number of steps required to obtain an independent sample is $O(\frac{\sigma_{max}}{\sigma_{min}})^2$.
- Generally, for a multivariate Gaussian, the number of steps required to obtain independent samples scales like $(\frac{\sigma_{max}}{\sigma_2})^2$ where σ_2 is the 2nd smallest standard deviation.
- If the length scales over which the distributions vary very different in different distribution, the Metropolis Hasting algorithm can have very slow convergence.

Comments

- Metropolis-Hasting Algorithm requires a proposal distribution for sampling. If the sample is high-dimensional, each time requires a sampling from a multi-variate distribution.
- Each time we obtain a sample.
- Is it possible to remove the requirement of proposal distribution? Or find a proposal distribution such that it matches the contour of the true distribution?
- Gibbs Sampling — Bingo!!

Gibbs Sampling

Suppose we have a distribution $p(z_1, z_2, z_3)$ over three variables. At step τ, we have selected values $z_1^{(\tau)}, z_2^{(\tau)}, z_3^{(\tau)}$.

1. Replace $z_1^{(\tau)}$ by a new value $z_1^{(\tau+1)}$ by sampling from

$$p(z_1 | z_2^{(\tau)}, z_3^{(\tau)})$$

2. Replace $z_2^{(\tau)}$ by a new value $z_2^{(\tau+1)}$ by sampling from

$$p(z_2 | z_1^{(\tau+1)}, z_3^{(\tau)})$$

3. Replace $z_3^{(\tau)}$ by a new value $z_3^{(\tau+1)}$ by sampling from

$$p(z_3 | z_1^{(\tau+1)}, z_2^{(\tau+1)})$$

4. Obtain a new sample $(z_1^{(\tau+1)}, z_2^{(\tau+1)}, z_3^{(\tau+1)})$. Repeat the above procedure.

Gibbs Sampling algorithm

Gibbs Sampling

1. Initialize $\{z_i : i = 1, \ldots, M\}$
2. For $\tau = 1, \ldots, T$:
   - Sample $z_1^{(\tau+1)} \sim p(z_1 | z_2^{(\tau)}, z_3^{(\tau)}, \ldots, z_M^{(\tau)})$.
   - Sample $z_2^{(\tau+1)} \sim p(z_2 | z_1^{(\tau+1)}, z_3^{(\tau)}, \ldots, z_M^{(\tau)})$.
   - $\vdots$
   - Sample $z_j^{(\tau+1)} \sim p(z_j | z_1^{(\tau+1)}, \ldots, z_{j-1}^{(\tau+1)}, z_{j+1}^{(\tau)}, \ldots, z_M^{(\tau)})$.
   - $\vdots$
   - Sample $z_M^{(\tau+1)} \sim p(z_M | z_1^{(\tau+1)}, z_2^{(\tau+1)}, \ldots, z_{M-1}^{(\tau+1)})$.

Note that each time we draw sample based on newly obtained value.

Invariant Distribution

- When we sample from $p(z_i|\{\mathbf{z}_{\setminus i}\})$, the marginal distribution $p(\mathbf{z}_{\setminus i})$ is unchanged as the values of $\mathbf{z}_{\setminus i}$ are not changed.
- Each step by definition, samples from the correct $p(z_i|\mathbf{z}_{\setminus i})$. So, the conditional and marginal distribution specify the joint distribution, which is invariant.
- To draw a new sample, multiple steps are performed. As each step is invariant, the distribution is invariant.

Ergodicity

- A sufficient condition: None of the conditional distribution be anywhere zero. In other words, any point in z space can be reached from any other point in a finite number of steps.
- If the above condition is not satisfied, need to analyze carefully.

Invariant Distribution

- When we sample from $p(z_i|\{\mathbf{z}_{\setminus i}\})$, the marginal distribution $p(\mathbf{z}_{\setminus i})$ is unchanged as the values of $\mathbf{z}_{\setminus i}$ are not changed.
- Each step by definition, samples from the correct $p(z_i|\mathbf{z}_{\setminus i})$. So, the conditional and marginal distribution specify the joint distribution, which is invariant.
- To draw a new sample, multiple steps are performed. As each step is invariant, the distribution is invariant.

Ergodicity

- A sufficient condition: None of the conditional distribution be anywhere zero. In other words, any point in z space can be reached from any other point in a finite number of steps.
- If the above condition is not satisfied, need to analyze carefully.

Connection to Metropolis-Hastings algorithm

Gibbs sampling can be considered as a special case of Metropolis-Hastings:

$$A(\mathbf{z}^\star, \mathbf{z}) = \frac{p(\mathbf{z}^\star) q_k(\mathbf{z}|\mathbf{z}^\star)}{p(\mathbf{z}) q_k(\mathbf{z}^\star|\mathbf{z})} = \frac{p(z_k^\star|\mathbf{z}_{\setminus k}^\star) p(\mathbf{z}_{\setminus k}^\star) p(z_k|\mathbf{z}_{\setminus k}^\star)}{p(z_k|\mathbf{z}_{\setminus k}) p(\mathbf{z}_{\setminus k}) p(z_k^\star|\mathbf{z}_{\setminus k})} = 1$$

where

$$q(z|z^*) = p(\mathbf{z}_{\setminus k}^*) p(z_k|\mathbf{z}_{\setminus k}^*)$$

and

$$z_{\setminus k}^* = z_{\setminus k}$$

Quick Review of Gaussian (1)

If

$$\begin{pmatrix} x_a \\ x_b \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}, \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}\right]$$

So

$$p(x_b|x_a) = N(\mu_b + \Sigma_{ba}\Sigma_{aa}^{-1}(x_a - \mu_a), \Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab})$$

Quick review of Gaussian (2)

If

$$\begin{aligned} P(x) &\sim N(\mu_x, \Sigma_x) \\ P(y|x) &\sim N(Ax + b, \Sigma_{y|x}) \end{aligned}$$

then

$$\begin{aligned} P(y) &\sim N(A\mu_x + b, \Sigma_{y|x} + A\Sigma_x A^T) \\ cov(x, y) &= \Sigma_x A^T \end{aligned}$$

Example

- Suppose the target distribution is

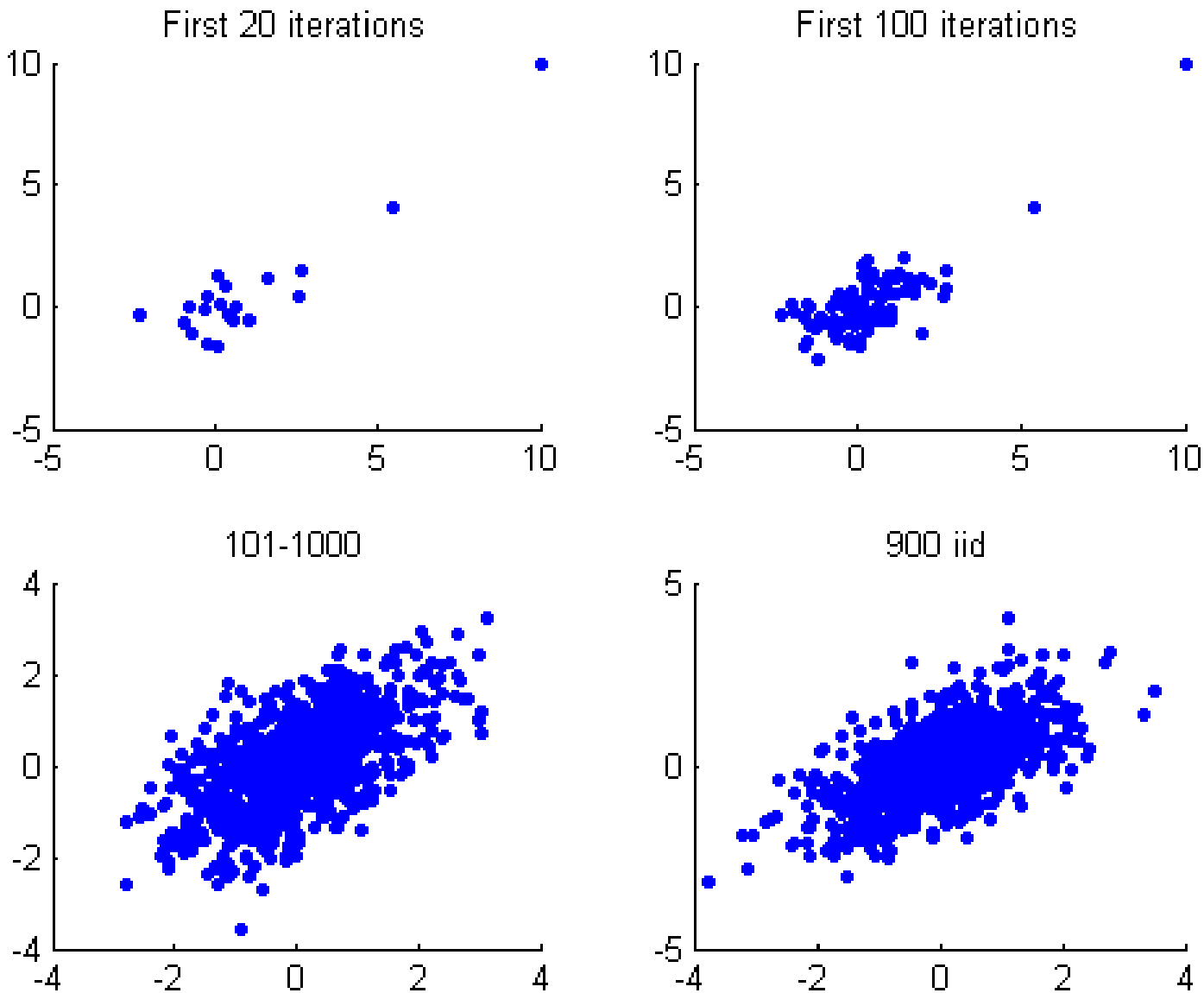
$$(X, Y) \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

- Gibbs sampler:

$$[X|Y = y] \sim N(\rho y, 1 - \rho^2)$$

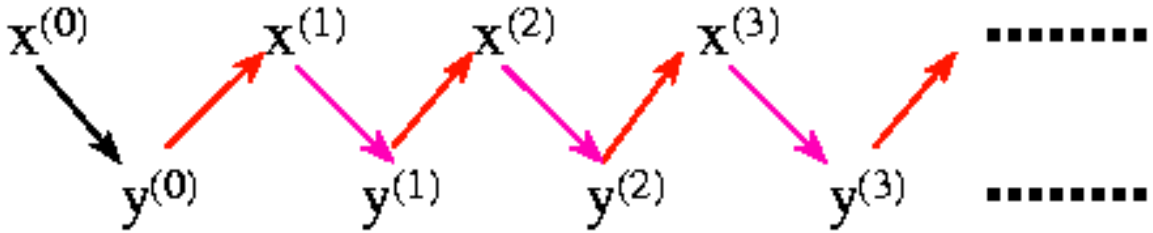
$$[Y|X = x] \sim N(\rho x, 1 - \rho^2)$$

Start from $(X, Y) = (10, 10)$, we can have a look at the trajectory


First 20 iterations
First 100 iterations
101-1000
900 iid

Why does it work?

- It is a *Markov Chain*!!



If $X^{(0)} = x_0$, then distribution of $X^{(t)}$ is $N(\rho^{2t} x_0, 1-\rho^{4t})$ which "converges" to $N(0,1)$ as $t \to \infty$.

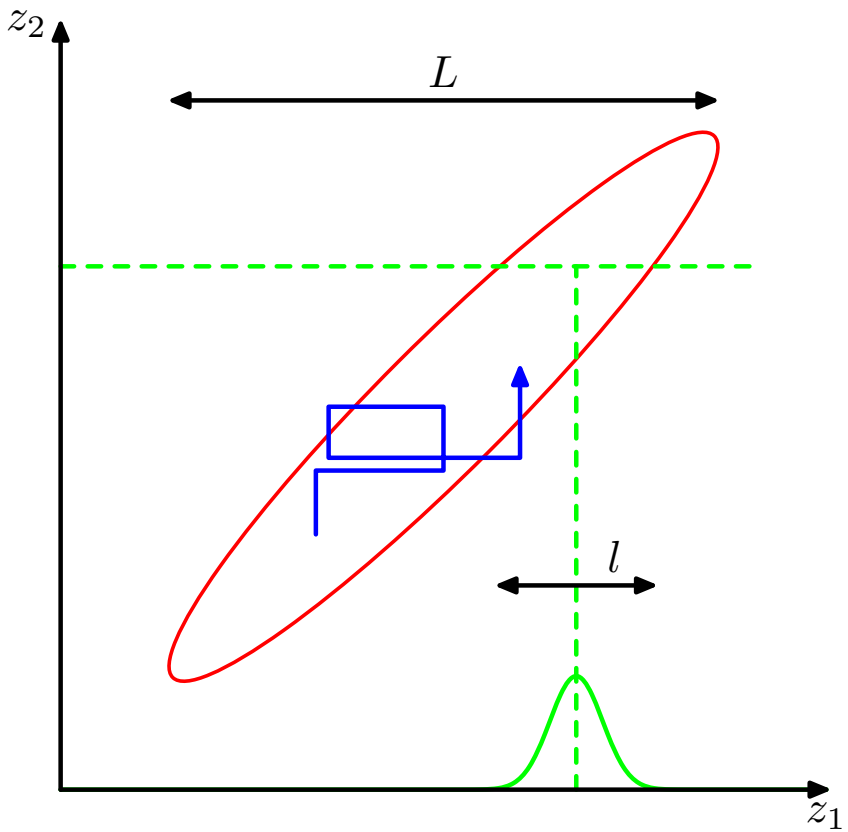
Joint distribution of $(X^{(t)}, Y^{(t)})$?

$$N\left(\begin{pmatrix}\rho^{2t}x_0\\ \rho^{2t+1}x_0\end{pmatrix},\begin{pmatrix}1-\rho^{4t} & \rho(1-\rho^{4t})\\ \rho(1-\rho^{4t}) & 1-\rho^{4t+2}\end{pmatrix}\right) \xrightarrow{t\to\infty} N\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right)$$

Random Walk Behavior

Marginal Distribution:

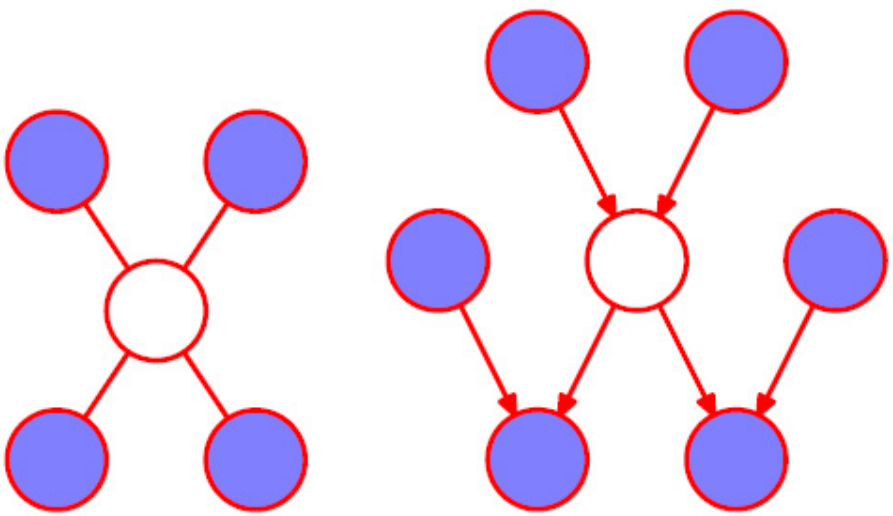
$$P(X) \sim N(0, 1);$$



- If ρ is large, then each step would move only a short distance ($\sigma^2 = 1 - \rho^2$). The number of steps to obtain independent samples would be of order $O((L/l)^2)$.
- If the Gaussian distribution is uncorrelated, then the Gibbs sampling procedure would be optimally efficient.
- Some approaches to reduce the random walk behavior in Gibbs sampling: *over-relaxation*.

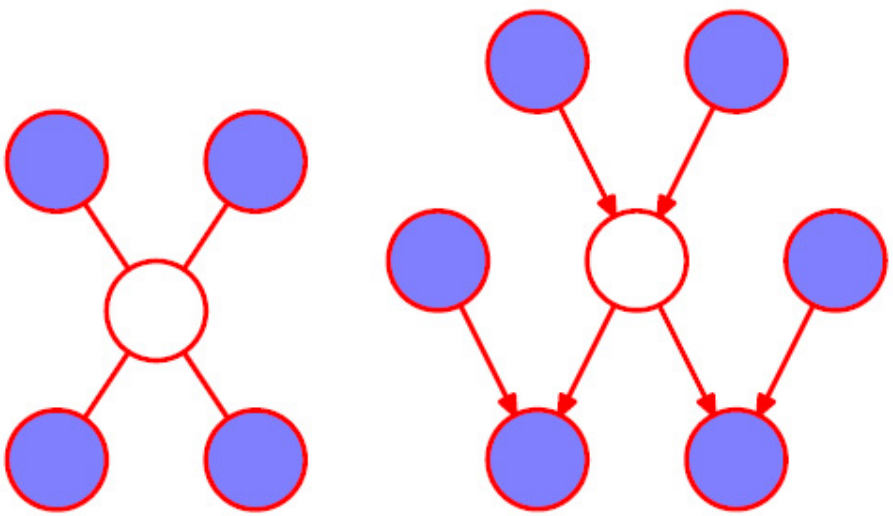
Applicability of Gibbs sampling

- Especially working on cases where joint distribution is unknown but conditional distribution is easy to sample.
- For undirected graphical models, only rely on its neighbors;
- For directed graphical models, only rely on its Markov Blanket.
- If the graph is constructed using exponential family distribution, and the parent-child relationship preserve conjugacy, then the conditional distribution in Gibbs sampling will have the same functional form.
- Usually, the full conditional distribution would be complicated. But if it's log concave, adaptive rejection sampling can be used.



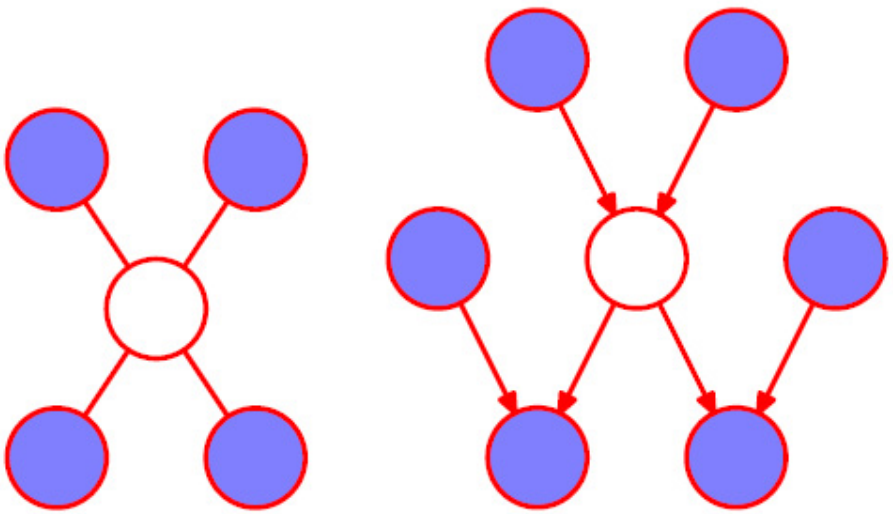
Applicability of Gibbs sampling

- Especially working on cases where joint distribution is unknown but conditional distribution is easy to sample.
- For undirected graphical models, only rely on its neighbors;
- For directed graphical models, only rely on its Markov Blanket.
- If the graph is constructed using exponential family distribution, and the parent-child relationship preserve conjugacy, then the conditional distribution in Gibbs sampling will have the same functional form.
- Usually, the full conditional distribution would be complicated. But if it's log concave, adaptive rejection sampling can be used.



Applicability of Gibbs sampling

- Especially working on cases where joint distribution is unknown but conditional distribution is easy to sample.
- For undirected graphical models, only rely on its neighbors;
- For directed graphical models, only rely on its Markov Blanket.
- If the graph is constructed using exponential family distribution, and the parent-child relationship preserve conjugacy, then the conditional distribution in Gibbs sampling will have the same functional form.
- Usually, the full conditional distribution would be complicated. But if it's log concave, adaptive rejection sampling can be used.



Factors to produce a good Monte Carlo estimate

- The amount of computation required for each transition;
- The time for the chain to converge to the equilibrium distribution; (Need to discard from the samples from the beginning)
- the number of transitions needed to move from one state drawn from the equilibrium distribution to another state that is almost independent. (which determines the number of states taken from the chain)