

# Statistical Models & Computing Methods

## Lecture 6&7: Markov Chain Monte Carlo



**Cheng Zhang**

School of Mathematical Sciences, Peking University

September 29 & October 04, 2022

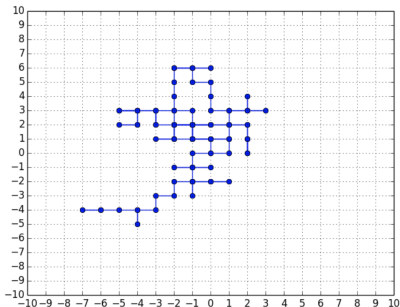
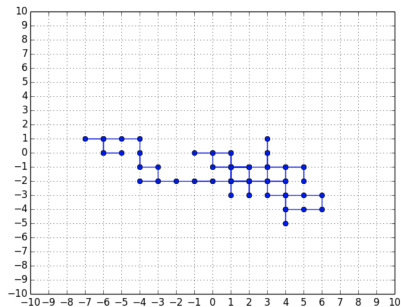
- ▶ Direct sampling in high-dimensional spaces is often **infeasible, very hard to get rare events**
- ▶ Rejection sampling, Importance sampling
  - ▶ Do not work well if the proposal  $q(x)$  is very different from  $f(x)$  or  $h(x)f(x)$ .
  - ▶ Moreover, **constructing appropriate  $q(x)$  can be difficult.** Making a good proposal usually requires knowledge of the analytic form of the target distribution - but if we had that, we wouldn't even need to sample
- ▶ Intuition: instead of a fixed proposal  $q(x)$ , what if we use an adaptive proposal?
- ▶ In this lecture, we are going to talk about one of the most popular sampling methods, **Markov chain Monte Carlo.**

- ▶ Stochastic processes is a family of random variables, usually indexed by a set of numbers (time). A discrete time stochastic process is simply a sequence of random variables,  $X_0, X_1, \dots, X_n$  defined on the same probability space
- ▶ One of the simplest stochastic processes (and one of the most useful) is the simple random walk
- ▶ Consider a simple random walk on a graph  $G = (\Omega, E)$ . The stochastic process starts from an initial position  $X_0 = x_0 \in \Omega$ , and proceeds following a simple rule:

$$p(X_{n+1}|X_n = x_n) \sim \text{Discrete}(\mathcal{N}(x_n)), \forall n \geq 0$$

where  $\mathcal{N}(x_n)$  denotes the neighborhood of  $x_n$

Two random walks on a  $20 \times 20$  grid graph



- ▶ The above simple random walk is a special case of another well-known stochastic process called *Markov chains*
- ▶ A Markov chain represents the stochastic movement of some particle in the state space over time. The particle initially starts from state  $i$  with probability  $\pi_i^{(0)}$ , and after that moves from the current state  $i$  at time  $t$  to the next state  $j$  with probability  $p_{ij}(t)$
- ▶ A Markov chain has three main elements:
  1. A state space  $\mathcal{S}$
  2. An initial distribution  $\pi^{(0)}$  over  $\mathcal{S}$
  3. Transition probabilities  $p_{ij}(t)$  which are non-negative numbers representing the probability of going from state  $i$  to  $j$ , and  $\sum_j p_{ij}(t) = 1$ .
- ▶ When  $p_{ij}(t)$  does not depend on time  $t$ , we say the Markov chain is time-homogenous



- ▶ Chain rule (in probability)

$$p(X_n = x_n, \dots, X_0 = x_0) = \prod_{i=1}^n p(X_i = x_i | X_{<i} = x_{<i})$$

- ▶ **Markov property**

$$p(X_{i+1} = x_{i+1} | X_i = x_i, \dots, X_0 = x_0) = p(X_{i+1} = x_{i+1} | X_i = x_i)$$

- ▶ Joint probability with Markov property

$$p(X_n = x_n, \dots, X_0 = x_0) = \prod_{i=1}^n p(X_i = x_i | X_{i-1} = x_{i-1})$$

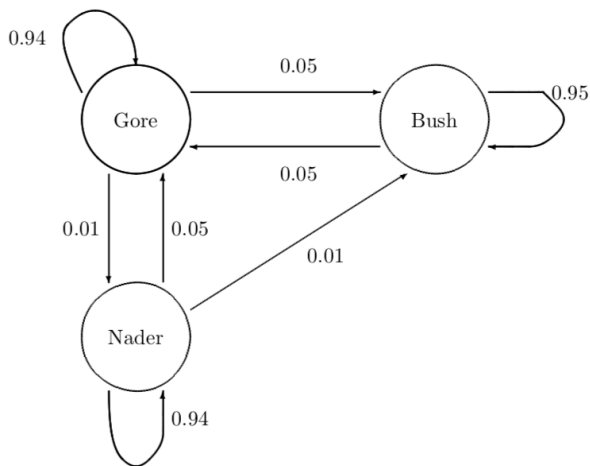
fully determined by the transition probabilities

- ▶ Consider the 2000 US presidential election with three candidates: Gore, Bush and Nader (just an illustrative example and does not reflect the reality of that election)
- ▶ We assume that the initial distribution of votes (i.e., probability of winning) was  $\pi = (0.49, 0.45, 0.06)$  for Gore, Bush and Nader respectively
- ▶ Further, we assume the following transition probability matrix

	<i>Gore</i>	<i>Bush</i>	<i>Nader</i>
<i>Gore</i>	0.94	0.05	0.01
<i>Bush</i>	0.05	0.95	0
<i>Nader</i>	0.05	0.01	0.94



A probabilistic graph presentation of the Markov chain





- ▶ If we represent the transition probability a square matrix  $P$  such that  $P_{ij} = p_{ij}$ , we can obtain the distribution of states in step  $n$ ,  $\pi^{(n)}$ , as follows

$$\pi^{(n)} = \pi^{(n-1)}P = \dots = \pi^{(0)}P^n$$

- ▶ For the above example, we have

$$\pi^{(0)} = (0.4900, 0.4500, 0.0600)$$

$$\pi^{(10)} = (0.4656, 0.4655, 0.0689)$$

$$\pi^{(100)} = (0.4545, 0.4697, 0.0758)$$

$$\pi^{(200)} = (0.4545, 0.4697, 0.0758)$$

- ▶ As we can see last, after several iterations, the above Markov chain converges to a distribution,  $(0.4545, 0.4697, 0.0758)$
- ▶ In this example, the chain would have reached this distribution regardless of what initial distribution  $\pi^{(0)}$  we chose. Therefore,  $\pi = (0.4545, 0.4697, 0.0758)$  is the stationary distribution for the above Markov chain
- ▶ **Stationary distribution.** A distribution of Markov chain states is called to be stationary if it remains the same in the next time step, i.e.,

$$\pi = \pi P$$

- ▶ How can we find out whether such distribution exists?
- ▶ Even if such distribution exists, is it unique or not?
- ▶ Also, how do we know whether the chain would converge to this distribution?
- ▶ To find out the answer, we briefly discuss some properties of Markov chains

- ▶ Irreducible: A Markov chain is **irreducible** if the chain can move from any state to another state.
- ▶ Examples
  - ▶ The simple random walk is irreducible
  - ▶ The following chain, however, is reducible since Nader does not communicate with the other two states (Gore and Bush)

	<i>Gore</i>	<i>Bush</i>	<i>Nader</i>
<i>Gore</i>	0.95	0.05	0
<i>Bush</i>	0.05	0.95	0
<i>Nader</i>	0	0	1



- ▶ Period: the period of a state  $i$  is the greatest common divisor of the times at which it is possible to move from  $i$  to  $i$ .
- ▶ For example, all the states in the following Markov chain have period 3.

$$\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

- ▶ Aperiodic: a Markov chain is said to be **aperiodic** if the period of each state is 1, otherwise the chain is periodic.



- ▶ **Recurrent** states: a state  $i$  is called recurrent if with probability 1, the chain would ever return to state  $i$  given that it started in state  $i$ .

	<i>Gore</i>	<i>Bush</i>	<i>Nader</i>
<i>Gore</i>	0.94	0.05	0.01
<i>Bush</i>	0.05	0.95	0
<i>Nader</i>	0.05	0.01	0.94

- ▶ **Positive recurrent:** a recurrent state  $j$  is called positive recurrent if the expected amount of time to return to state  $j$  given that the chain started in state  $j$  is finite
- ▶ For a positive recurrent Markov chain, the stationary distribution exists and is unique



- ▶ **Reversibility:** a Markov chain is said to be reversible with respect to a probability distribution  $\pi$  if  $\pi_i p_{ij} = \pi_j p_{ji}$
- ▶ In fact, if a Markov chain is reversible with respect to  $\pi$ , then  $\pi$  is also a stationary distribution

$$\begin{aligned}\sum_i \pi_i p_{ij} &= \sum_i \pi_j p_{ji} \\ &= \pi_j \sum_i p_{ji} \\ &= \pi_j\end{aligned}$$

since  $\sum_i p_{ji} = 1$  for all transition probability matrices

- ▶ This is also known as *detailed balance condition*

- ▶ We can define a Markov chain on a general state space  $\mathcal{X}$  with initial distribution  $\pi^{(0)}$  and transition probabilities  $p(x, A)$  defined as the probability of jumping to the subset  $A$  from point  $x \in \mathcal{X}$
- ▶ Similarly, with Markov property, we have the joint probability

$$p(X_0 \in A_0, \dots, X_n \in A_n) = \int_{A_0} \pi^{(0)}(dx_0) \dots \int_{A_n} p(x_{n-1}, dx_n)$$

- ▶ Example. Consider a Markov chain with the real line as its state space. The initial distribution is  $\mathcal{N}(0, 1)$ , and the transition probability is  $p(x, \cdot) = \mathcal{N}(x, 1)$ . This is just a **Brownian motion** (observed at discrete time)





- ▶ Unlike the discrete space, we now need to talk about the property of Markov chains with a continuous non-zero measure  $\phi$ , on  $\mathcal{X}$ , and use sets  $A$  instead of points
- ▶ A chain is  $\phi$ -irreducible if for all  $A \subseteq \mathcal{X}$  with  $\phi(A) > 0$  and for all  $x \in \mathcal{X}$ , there exists a positive integer  $n$  such that

$$p^n(x, A) = p(X_n \in A | X_0 = x) > 0$$

- ▶ Similarly, we need to modify our definition of period

- ▶ A distribution  $\pi$  is a stationary distribution if

$$\pi(A) = \int_{\mathcal{X}} \pi(dx)p(x, A), \quad \forall A \subseteq \mathcal{X}$$

- ▶ As for the discrete case, a continuous space Markov chain is reversible with respect to  $\pi$  if

$$\pi(dx)p(x, dy) = \pi(dy)p(y, dx)$$

- ▶ Similarly, if the chain is reversible with respect to  $\pi$ , then  $\pi$  is a stationary distribution
- ▶ Example. Consider a Markov chain on the real line with initial distribution  $\mathcal{N}(1, 1)$  and transition probability  $p(x, \cdot) = \mathcal{N}(\frac{x}{2}, \frac{3}{4})$ . It is easy to show that the chain converges to  $\mathcal{N}(0, 1)$  (**Exercise**)



- ▶ Ergodic: a Markov chain is ergodic if it is both irreducible and aperiodic, with stationary distribution  $\pi$
- ▶ **Ergodic Theorem.** For an ergodic Markov chain on the state space  $\mathcal{X}$  having stationary distribution  $\pi$ , we have: (i) for all measurable  $A \subseteq \mathcal{X}$  and  $\pi$ -a.e.  $x \in \mathcal{X}$ ,

$$\lim_{t \rightarrow \infty} p^t(x, A) = \pi(A)$$

(ii)  $\forall f$  with  $\mathbb{E}_\pi |f(x)| < \infty$ ,

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T f(X_t) = \int_{\mathcal{X}} f(x) \pi(x) dx, \quad \text{a.s.}$$

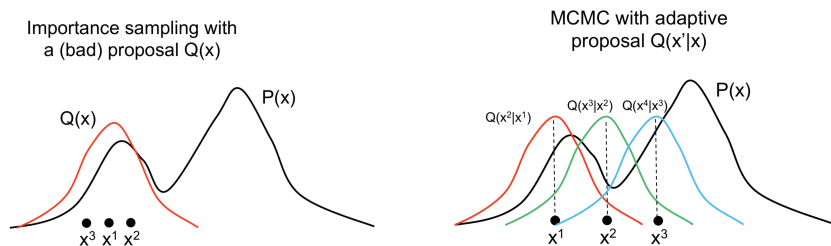
In particular,  $\pi$  is the **unique** stationary probability density function for the chain



- ▶ Now suppose we are interested in sampling from a distribution  $\pi$  (e.g., the unnormalized posterior)
- ▶ Markov chain Monte Carlo (MCMC) is a method that samples from a Markov chain whose stationary distribution is the target distribution  $\pi$ . It does this by constructing an appropriate transition probability for  $\pi$
- ▶ MCMC, therefore, can be viewed as an **inverse** process of Markov chains



- ▶ The transition probability in MCMC resembles the proposal distribution we used in previous Monte Carlo methods.
- ▶ Instead of using a fixed proposal (as in importance sampling and rejection sampling), MCMC algorithms feature **adaptive proposals**



Figures adapted from Eric Xing (CMU)



- ▶ Suppose that we are interested in sampling from a distribution  $\pi$ , whose density we know up to a constant  $P(x) \propto \pi(x)$
- ▶ We can construct a Markov chain with a transition probability (i.e., proposal distribution)  $Q(x'|x)$  which is symmetric; that is,  $Q(x'|x) = Q(x|x')$
- ▶ Example. A normal distribution with the mean at the current state and fixed variance  $\sigma^2$  is symmetric since

$$\exp\left(-\frac{(y-x)^2}{2\sigma^2}\right) = \exp\left(-\frac{(x-y)^2}{2\sigma^2}\right)$$



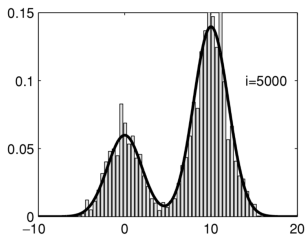
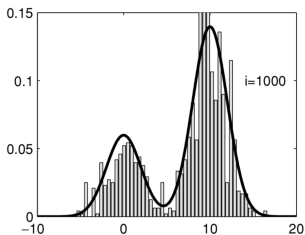
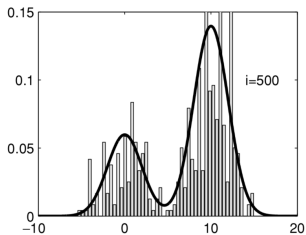
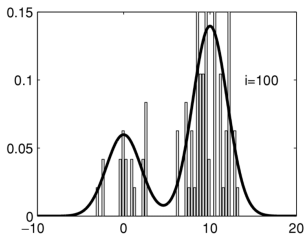
In each iteration we do the following

- ▶ Draws a sample  $x'$  from  $Q(x'|x)$ , where  $x$  is the previous sample
- ▶ Calculated the acceptance probability

$$a(x'|x) = \min \left( 1, \frac{P(x')}{P(x)} \right)$$

Note that we only need to compute  $\frac{P(x')}{P(x)}$ , the unknown constant cancels out

- ▶ Accept the new sample with probability  $a(x'|x)$  or remain at state  $x$ . The acceptance probability ensures that, after sufficient many draws, our samples will come from the true distribution  $\pi(x)$



Adapted from Andrieu, Freitas, Doucet, Jordan, 2003





- ▶ How do we know that the chain is going to converge to  $\pi$ ?
- ▶ Suppose the support of the proposal distribution is  $\mathcal{X}$  (e.g., Gaussian distribution), then the Markov chain is irreducible and aperiodic.
- ▶ We only need to verify the detailed balance condition

$$\begin{aligned}\pi(dx)p(x, dx') &= \pi(x)dx \cdot Q(x'|x)a(x'|x)dx' \\ &= \pi(x)Q(x'|x) \min\left(1, \frac{\pi(x')}{\pi(x)}\right) dx dx' \\ &= Q(x'|x) \min(\pi(x), \pi(x')) dx dx' \\ &= Q(x|x') \min(\pi(x'), \pi(x)) dx dx' \\ &= \pi(x')dx' \cdot Q(x|x') \min\left(1, \frac{\pi(x)}{\pi(x')}\right) dx \\ &= \pi(dx')p(x', dx)\end{aligned}$$



- ▶ It turned out that symmetric proposal distribution is not necessary. Hastings (1970) later on generalized the above algorithm using the following acceptance probability for general  $Q(x'|x)$

$$a(x'|x) = \min \left( 1, \frac{P(x')Q(x|x')}{P(x)Q(x'|x)} \right)$$

- ▶ Similarly, we can show that detailed balanced condition is preserved

- ▶ Under mild assumptions on the proposal distribution  $Q$ , the algorithm is ergodic
- ▶ However, the choice of  $Q$  is important since it determines the speed of convergence to  $\pi$  and the efficiency of sampling
- ▶ Usually, the proposal distribution depend on the current state. But it can be independent of current state, which leads to an independent MCMC sampler that is somewhat like a rejection/importance sampling method
- ▶ Some examples of commonly used proposal distributions
  - ▶  $Q(x'|x) \sim \mathcal{N}(x, \sigma^2)$
  - ▶  $Q(x'|x) \sim \text{Uniform}(x - \delta, x + \delta)$
- ▶ Finding a good proposal distribution is hard in general

- ▶ Recall the univariate Gaussian model with known variance

$$y_i \sim \mathcal{N}(\theta, \sigma^2)$$
$$p(y|\theta, \sigma) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \theta)^2}{2\sigma^2}\right)$$

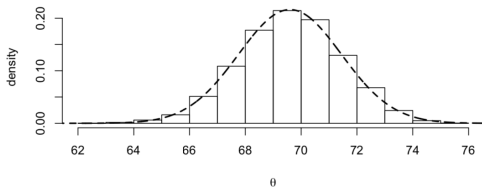
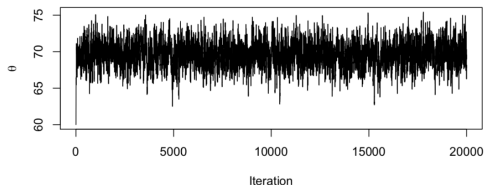
- ▶ Note that there is a conjugate  $\mathcal{N}(\mu_0, \tau_0^2)$  prior for  $\theta$ , and the posterior has a close form normal distribution
- ▶ Now let's pretend that we don't know this exact posterior distribution and use a Markov chain to sample from it.

- ▶ We can of course write the posterior distribution up to a constant

$$p(\theta|y) \propto \exp\left(-\frac{(\theta - \mu_0)^2}{2\tau_0^2}\right) \prod_{i=1}^n \exp\left(-\frac{(y_i - \theta)^2}{2\sigma^2}\right) = P(\theta)$$

- ▶ We use  $\mathcal{N}(\theta^{(i)}, 1)$ , a normal distribution around our current state, to propose the next step
- ▶ Starting from an initial point  $\theta^{(0)}$  and propose the next step  $\theta' \sim \mathcal{N}(\theta^{(0)}, 1)$ , we either accept this value with probability  $\alpha(\theta'|\theta^{(0)})$  or reject and stay where we are
- ▶ We continue these steps for many iterations

- ▶ As we can see, the posterior distribution we obtained using the Metropolis algorithm is very similar to the exact posterior



- ▶ Now suppose we want to model the number of half court shots Stephen Curry has made in a game using Poisson model

$$y_i \sim \text{Poisson}(\theta)$$

- ▶ He made 0 and 1 half court shots in the first two games respectively
- ▶ We used  $\text{Gamma}(1.4, 10)$  prior for  $\theta$ , and because of conjugacy, the posterior distribution also had a Gamma distribution

$$\theta|y \sim \text{Gamma}(2.4, 12)$$

- ▶ Again, let's ignore the closed form posterior and use MCMC for sampling the posterior distribution

- ▶ The prior is

$$p(\theta) \propto \theta^{0.4} \exp(-10\theta)$$

- ▶ The likelihood is

$$p(y|\theta) \propto \theta^{y_1+y_2} \exp(-2\theta)$$

where  $y_1 = 0$  and  $y_2 = 1$

- ▶ Therefore, the posterior is proportional to

$$p(\theta|y) \propto \theta^{0.4} \exp(-10\theta) \cdot \theta^{y_1+y_2} \exp(-2\theta) = P(\theta)$$





- ▶ Symmetric proposal distributions such as

$$\text{Uniform}(\theta^{(i)} - \delta, \theta^{(i)} + \delta) \text{ or } \mathcal{N}(\theta^{(i)}, \sigma^2)$$

might not be efficient since they do not take the non-negative support of the posterior into account.

- ▶ Here, we use a non-symmetric proposal distribution such as  $\text{Uniform}(0, \theta^{(i)} + \delta)$  and use the Metropolis-Hastings (MH) algorithm instead
- ▶ We set  $\delta = 1$

We start from  $\theta_0 = 1$  and follow these steps in each iteration

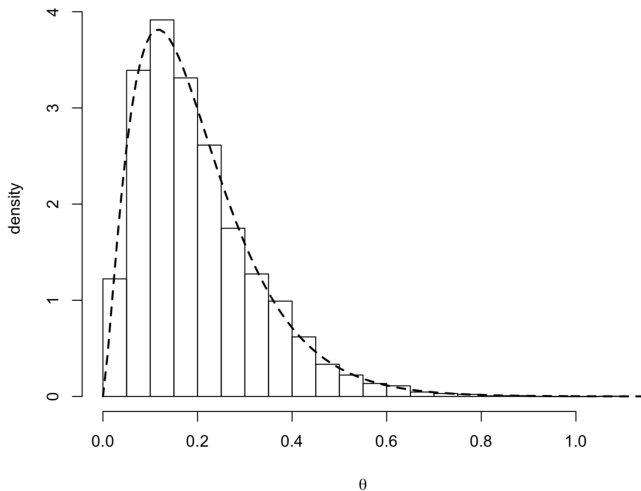
- ▶ Sample  $\theta'$  from  $\mathcal{U}(0, \theta^{(i)} + 1)$
- ▶ Calculate the acceptance probability

$$a(\theta'|\theta^{(i)}) = \min \left( 1, \frac{P(\theta')\text{Uniform}(\theta^{(i)}|0, \theta' + 1)}{P(\theta^{(i)})\text{Uniform}(\theta'|0, \theta^{(i)} + 1)} \right)$$

- ▶ Sample  $u \sim \mathcal{U}(0, 1)$  and set

$$\theta^{(i+1)} = \begin{cases} \theta' & u < a(\theta'|\theta^{(i)}) \\ \theta^{(i)} & \text{otherwise} \end{cases}$$





- ▶ What if the distribution is multidimensional, *i.e.*,  
 $x = (x_1, x_2, \dots, x_d)$
- ▶ We can still use the Metropolis algorithm (or MH), with a multivariate proposal distribution, *i.e.*, we now propose  
 $x' = (x'_1, x'_2, \dots, x'_d)$
- ▶ For example, we can use a multivariate normal  $\mathcal{N}_d(x, \sigma^2 I)$ , or a  $d$ -dimensional uniform distribution around the current state

- ▶ Here we construct a banana-shaped posterior distribution as follows

$$y|\theta \sim \mathcal{N}(\theta_1 + \theta_2^2, \sigma_y^2), \quad \sigma_y = 2$$

We generate data  $y_i \sim \mathcal{N}(1, \sigma_y^2)$

- ▶ We use a bivariate normal prior for  $\theta$

$$\theta = (\theta_1, \theta_2) \sim \mathcal{N}(0, I)$$

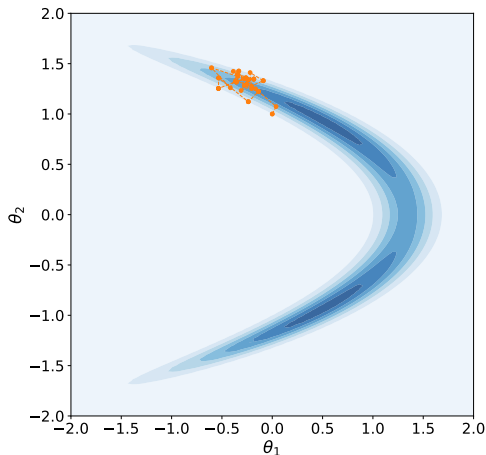
- ▶ The posterior is

$$p(\theta|y) \propto \exp\left(-\frac{\theta_1^2 + \theta_2^2}{2}\right) \cdot \exp\left(-\frac{\sum_i (y_i - \theta_1 - \theta_2^2)^2}{2\sigma_y^2}\right)$$

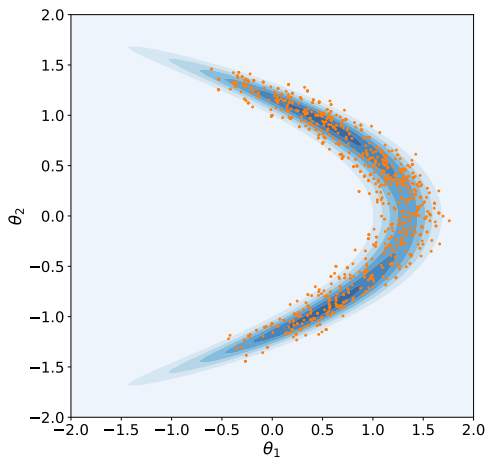
- ▶ We use the Metropolis algorithm to sample from posterior, with a bivariate normal proposal distribution such as  $\mathcal{N}(\theta^{(i)}, (0.15)^2 I)$



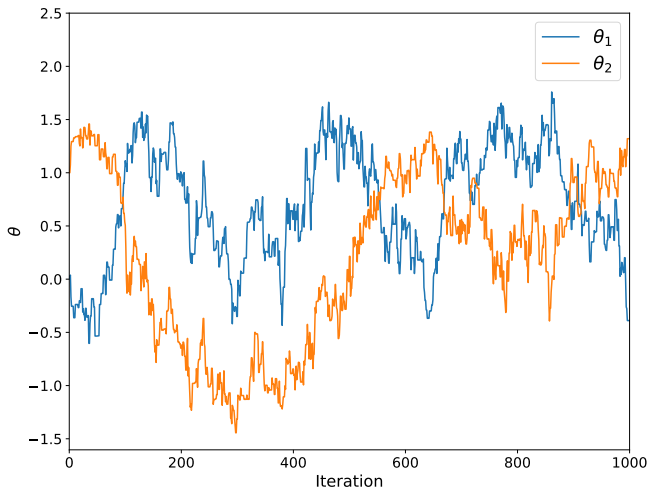
The first few samples from the posterior distribution of  $\theta = (\theta_1, \theta_2)$ , using a bivariate normal proposal



Posterior samples for  $\theta = (\theta_1, \theta_2)$



Trace plot of posterior samples for  $\theta = (\theta_1, \theta_2)$





- ▶ Sometimes, it is easier to decompose the parameter space into several components, and use the Metropolis (or MH) algorithm for one component at a time
- ▶ At iteration  $i$ , given the current state  $(x_1^{(i)}, \dots, x_d^{(i)})$ , we do the following for all components  $k = 1, 2, \dots, d$ 
  - ▶ Sample  $x'_k$  from the univariate proposal distribution  $Q(x'_k | \dots, x_{k-1}^{(i+1)}, x_k^{(i)}, \dots)$
  - ▶ Accept this new value and set  $x_k^{(i+1)} = x'_k$  with probability

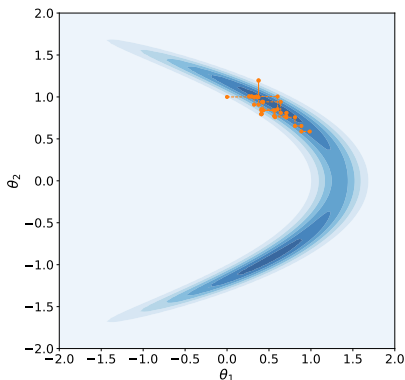
$$a(x'_k | \dots, x_{k-1}^{(i+1)}, x_k^{(i)}, \dots) = \min \left( 1, \frac{P(\dots, x_{k-1}^{(i+1)}, x'_k, \dots)}{P(\dots, x_{k-1}^{(i+1)}, x_k^{(i)}, \dots)} \right)$$

or reject it and set  $x_k^{(i+1)} = x_k^{(i)}$



- ▶ Note that in general, we can decompose the space of random variable into blocks of components
- ▶ Also, we can update the components sequentially or randomly
- ▶ As long as each transition probability individually leaves the target distribution invariant, their sequence would leave the target distribution invariant
- ▶ In Bayesian models, this is especially useful if it is easier and computationally less intensive to evaluate the posterior distribution when one subset of parameters change at a time

- ▶ In the example of banana-shaped distribution, we can sample  $\theta_1$  and  $\theta_2$  one at a time
- ▶ The first few samples from the posterior distribution of  $\theta = (\theta_1, \theta_2)$ , using a univariate normal proposal sequentially



- ▶ As the dimensionality of the parameter space increases, it becomes difficult to find an appropriate proposal distributions (e.g., with appropriate step size) for the Metropolis (or MH) algorithm
- ▶ If we are lucky (in some situations we are!), the conditional distribution of one component,  $x_j$ , given all other components,  $x_{-j}$  is tractable and has a close form so that we can sample from it directly
- ▶ If that's the case, we can sample from each component one at a time using their corresponding conditional distributions  $P(x_j|x_{-j})$

- ▶ This is known as the Gibbs sampler (GS) or “heat bath” (Geman and Geman, 1984)
- ▶ Note that in Bayesian analysis, we are mainly interested in sampling from  $p(\theta|y)$
- ▶ Therefore, we use the Gibbs sampler when  $P(\theta_j|y, \theta_{-j})$  has a closed form, e.g., there is a conditional conjugacy
- ▶ One example is the univariate normal model. As we will see later, given  $\sigma$ , the posterior  $P(\mu|y, \sigma^2)$  has a closed form, and given  $\mu$ , the posterior distribution of  $P(\sigma^2|\mu, y)$  also has a closed form

- ▶ The Gibbs sampler works as follows
- ▶ Initialize starting value for  $x_1, x_2, \dots, x_d$
- ▶ At each iteration, pick an ordering of the  $d$  variables (can be sequential or random)
  1. Sample  $x \sim P(x_i | x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_d)$ , *i.e.*, the conditional distribution of  $x_i$  given the current values of all other variables
  2. Update  $x_i \leftarrow x$
- ▶ When we update  $x_i$ , we immediately use its new value for sampling other variables  $x_j$

- ▶ Note that in GS, we are not proposing anymore, we are directly sampling, which can be viewed as a proposal that will **always be accepted**
- ▶ This way, the Gibbs sampler can be viewed as a special case of MH, whose proposal is

$$Q(x'_i, x_{-i} | x_i, x_{-i}) = P(x'_i | x_{-i})$$

- ▶ Applying MH with this proposal, we obtain

$$\begin{aligned} a(x'_i, x_{-i} | x_i, x_{-i}) &= \min \left( 1, \frac{P(x'_i, x_{-i})Q(x_i, x_{-i} | x'_i, x_{-i})}{P(x_i, x_{-i})Q(x'_i, x_{-i} | x_i, x_{-i})} \right) \\ &= \min \left( 1, \frac{P(x'_i, x_{-i})P(x_i | x_{-i})}{P(x_i, x_{-i})P(x'_i | x_{-i})} \right) = \min \left( 1, \frac{P(x'_i, x_{-i})P(x_i, x_{-i})}{P(x_i, x_{-i})P(x'_i, x_{-i})} \right) \\ &= 1 \end{aligned}$$



- ▶ We can now use the Gibbs sampler to simulate samples from the posterior distribution of the parameters of a univariate normal  $y \sim \mathcal{N}(\mu, \sigma^2)$  model, with prior

$$\mu \sim \mathcal{N}(\mu_0, \tau_0^2), \quad \sigma^2 \sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2)$$

- ▶ Given  $(\sigma^{(i)})^2$  at the  $i^{\text{th}}$  iteration, we sample  $\mu^{(i+1)}$  from

$$\mu^{(i+1)} \sim \mathcal{N}\left(\frac{\frac{\mu_0}{\tau_0^2} + \frac{n\bar{y}}{(\sigma^{(i)})^2}}{\frac{1}{\tau_0^2} + \frac{n}{(\sigma^{(i)})^2}}, \frac{1}{\frac{1}{\tau_0^2} + \frac{n}{(\sigma^{(i)})^2}}\right)$$

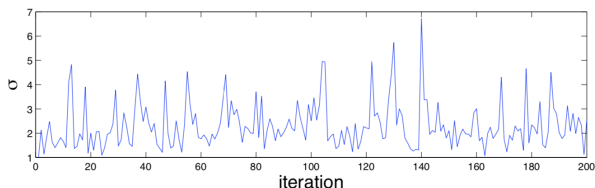
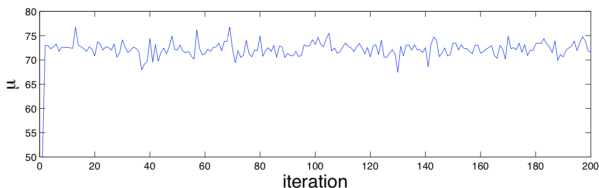
- ▶ Given  $\mu^{(i+1)}$ , we sample a new  $\sigma^2$  from

$$(\sigma^{(i+1)})^2 \sim \text{Inv-}\chi^2\left(\nu_0 + n, \frac{\nu_0 \sigma_0^2 + \nu n}{\nu_0 + n}\right), \quad \nu = \frac{1}{n} \sum_{j=1}^n (y_j - \mu^{(i+1)})^2$$





- ▶ The following graphs show the trace plots of the posterior samples (for both  $\mu$  and  $\sigma$ )

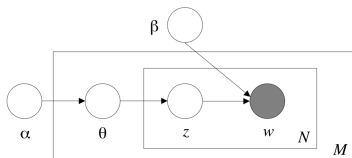


Gibbs sampling algorithms have been widely used in **probabilistic graphical models**

- ▶ Conditional distributions are fairly easy to derive for many graphical models (e.g., mixture models, Latent Dirichlet allocation)
- ▶ Have reasonable computation and memory requirements, only needs to sample one random variable at a time
- ▶ Can be Rao-Blackwellized (integrate out some random variable) to decrease the sampling variance. This is called *collapsed Gibbs sampling*



- ▶ Generative model of documents (Blei, Jordan and Ng, 2003). Also broadly applicable to collaborative filtering, image retrieval, bioinformatics, etc.

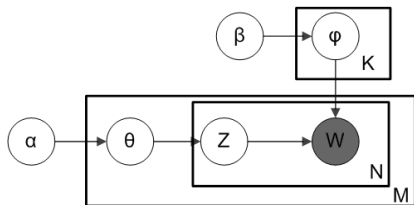


- ▶ choose a mixture of topics the document:  $\theta \sim \text{Dir}(\alpha)$
- ▶ choose a topic for each of the document:

$$z_n \sim \text{Multinomial}(\theta)$$

- ▶ choose word given the topic:  $w_n | z_n, \beta \sim p(w_n | z_n, \beta)$





- ▶ Use the probability model for LDA, with an additional Dirichlet prior on  $\phi$ .
- ▶ The complete probability model

$$\begin{aligned}
 w_i | z_i, \phi^{(z_i)} &\sim \text{Discrete}(\phi^{(z_i)}) \\
 \phi &\sim \text{Dirichlet}(\beta) \\
 z_i | \theta^{(d_i)} &\sim \text{Discrete}(\theta^{(d_i)}) \\
 \theta &\sim \text{Dirichlet}(\alpha)
 \end{aligned}$$



- ▶ The joint probability is

$$p(w, z, \phi, \theta | \alpha, \beta) = \prod_i p(w_i | z_i, \phi^{(z_i)}) p(\phi | \beta) \cdot \prod_i p(z_i | \theta^{(d_i)}) p(\theta | \alpha)$$

- ▶ Due to conjugate priors, we can easily integrate out  $\phi$  and  $\theta$  (T. Griffiths & M. Steyvers, 2004)

$$p(w|z) = \left( \frac{\Gamma(V\beta)}{\Gamma(\beta)^V} \right)^K \prod_{j=1}^K \frac{\prod_w \Gamma(n_j^{(w)} + \beta)}{\Gamma(n_j^{(\cdot)} + V\beta)}$$

$$p(z) = \left( \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \right)^M \prod_{d=1}^M \frac{\prod_j \Gamma(n_j^{(d)} + \alpha)}{\Gamma(n^{(d)} + K\alpha)}$$

$n_j^{(w)}$  ← number of times word  $w$  assigned to topic  $j$

$n_j^{(d)}$  ← number of times topic  $j$  used in document  $d$




- ▶ Need full conditional distributions for variables
- ▶ We only sample  $z$ , whose conditional distributions is

$$p(z_i = j | z_{-i}, w) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + V\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + K\alpha}$$



- ▶ Need full conditional distributions for variables
- ▶ We only sample  $z$ , whose conditional distributions is

$$p(z_i = j | z_{-i}, w) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + V\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + K\alpha}$$



probability of  $w_i$  under topic  $j$



- ▶ Need full conditional distributions for variables
- ▶ We only sample  $z$ , whose conditional distributions is

$$p(z_i = j | z_{-i}, w) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + V\beta}$$

probability of  $w_i$  under topic  $j$ 
probability of topic  $j$  in document  $d_i$





- ▶ Need full conditional distributions for variables
- ▶ We only sample  $z$ , whose conditional distributions is

$$p(z_i = j | z_{-i}, w) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + V\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + K\alpha}$$

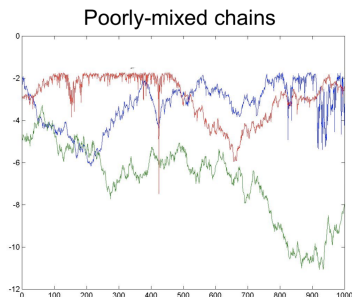
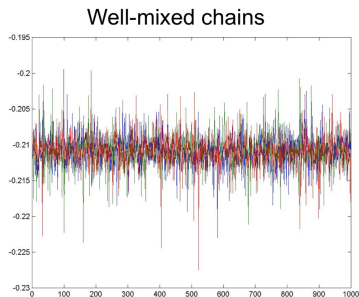
probability of  $w_i$  under topic  $j$ 
probability of topic  $j$  in document  $d_i$

- ▶ This is nicer than your average Gibbs sampler:
  - ▶ memory: counts can be cached in two sparse matrices
  - ▶ the distributions on  $\phi$  and  $\theta$  are analytic given  $z$  and  $w$ , and can later be found for each sample



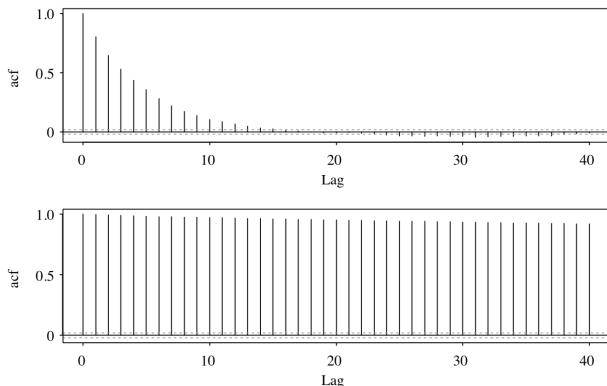
- ▶ For more complex models, we might only have conditional conjugacy for one part of the parameters
- ▶ In such situations, we can combine the Gibbs sampler with the Metropolis method
- ▶ That is, we update the components with conditional conjugacy using Gibbs sampler and for the rest parameters, we use the Metropolis (or MH)

- ▶ MCMC would converge to the target distribution if run sufficiently long
- ▶ However, it is often non-trivial to determine whether the chain has converged or not in practice
- ▶ Also, how do we measure the efficiency of MCMC chains?
- ▶ In what follows, we will discuss some practical advice for coding MCMC algorithms



Monitor convergence by plotting samples from multiple MH runs (chains)

- ▶ If the chains are well-mixed (left), they are probably converged
- ▶ If the chains are poorly-mixed (right), we may need to continue burn-in



- ▶ An autocorrelation plot summarizes the correlation in the sequence of a Markov chain at different iteration lags
- ▶ A chain that has poor mixing will exhibit slow decay of the autocorrelation as the lag increases



- ▶ Since MCMC samples are correlated, *effective sample size* are often used to measure the efficiency when MCMC samples are used for estimation instead of independent samples
- ▶ The effective sample size (ESS) is defined as

$$\text{ESS} = \frac{n}{1 + 2 \sum_{k=1}^{\infty} \rho(k)}$$

where  $\rho(k)$  is the autocorrelation at lag  $k$

- ▶ ESS are commonly used to compare the efficiency of competing MCMC samplers for a given problem. Larger ESS usually means faster convergence

- ▶ One of the hardest problem to diagnose is whether or not the chain has become stuck in one or more modes of the target distribution
- ▶ In this case, all convergence diagnostics may indicate that the chain has converged, though it does not
- ▶ A partial solution: run multiple chains and compare the within- and between-chain behavior

- ▶ Auxiliary variable strategies can be used to improving mixing of Markov chains
- ▶ When standard MCMC methods mix poorly, one potential remedy is to augment the state space of the variable of interest
- ▶ This approach can lead to chains that mix faster and require less tuning than the standard MCMC methods
- ▶ Main idea: construct a Markov chain over  $(X, U)$  ( $U$  is the auxiliary variable) with stationary distribution marginalizes to the target distribution of  $X$
- ▶ As we will see later, this includes a large family of modern MCMC methods



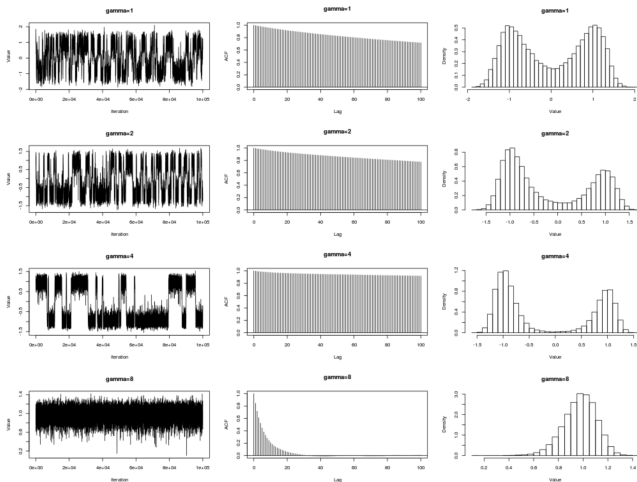
- ▶ Suppose that we have a challenging target distribution  $f(x) \propto \exp(-U(x))$
- ▶ We can introduce temperatures to construct a sequence of distributions that are easier to sample from

$$f_k(x) \propto \exp(-U(x)/T_k), \quad k = 0, \dots, K$$

where  $1 = T_0 < T_1 < \dots < T_K$

- ▶ When simulating Markov chains with different temperature  $T$ , the chain with high temperature (hot chain) is likely to mix better than the chain with cold temperature (cold chain)
- ▶ Therefore, we can run parallel chains and swap states between the chains to improve mixing

$$f_T(x) \propto \exp(-(x^2 - 1)^2/T), \quad T = 1/\gamma$$



We run parallel Markov chains for distributions with different temperatures. In each iteration

- ▶ Follow regular Metropolis steps in each chain to get new states  $x_0^{(t)}, \dots, x_K^{(t)}$
- ▶ Select two temperatures, say  $(i, j), i < j$ , and swap the states

$$x_0^{(t)}, \dots, x_i^{(t)}, \dots, x_j^{(t)}, \dots, x_K^{(t)} \rightarrow x_0^{(t)}, \dots, x_j^{(t)}, \dots, x_i^{(t)}, \dots, x_K^{(t)}$$

- ▶ Accept the swapped new states with the following probability

$$\min \left( 1, f_i(x_j^{(t)}) f_j(x_i^{(t)}) / f_i(x_i^{(t)}) f_j(x_j^{(t)}) \right)$$



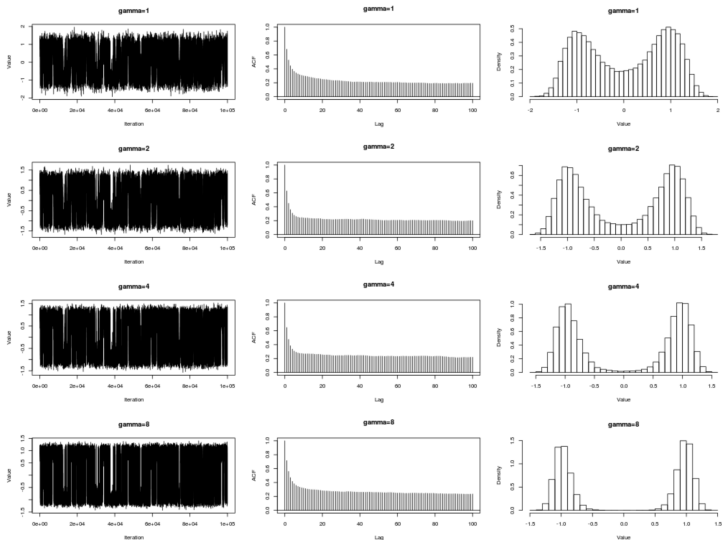
- ▶ Both the within-chain Metropolis updates and the between-chain swap preserves

$$p(x_0, \dots, x_K) \propto f_0(x_0) f_1(x_1) \dots f_K(x_K)$$

- ▶ Therefore, the joint distribution of  $(x_0^{(t)}, \dots, x_K^{(t)})$  will converge to  $p(x)$ , and the marginal distribution of  $x_0$  (cold chain) is the target distribution
- ▶ There are many ways to swap chains. For example, we can pick a pair of temperatures uniformly at random or only swap chains with successive temperatures
- ▶ The design of temperature levels could be crucial for the performance

# Example: Double-well Potential Distribution

66/75



- ▶ Slice sampling was introduced by Neal (2003) to accelerate mixing of Metropolis (or MH)
- ▶ It is essentially a Gibbs sampler in the augmented space  $(X, U)$  with density

$$f(x, u) = f(x)f(u|x)$$

where  $U$  is the auxiliary variable and  $f(u|x)$  is designed to be a uniform distribution  $\mathcal{U}(0, f(x))$

- ▶ For this purpose, slice sampling alternates between two steps:
  - ▶ Given the current state of the Markov chain,  $x$ , we uniformly sample a new point  $u$  from the interval  $(0, f(x))$

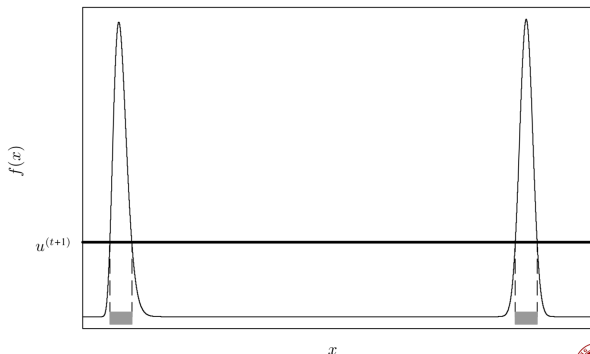
$$U|x \sim \mathcal{U}(0, f(x))$$

- ▶ Given the current value of  $u$ , we uniformly sample from the region  $S = \{x : f(x) > u\}$ , which is referred to as the *slice* defined by  $u$

$$X|u \sim \mathcal{U}(S)$$

- ▶ As mentioned by Neal (2003), in practice it is safer to compute  $g(x) = \log(f(x))$ , and use the auxiliary variable  $z = \log(u) = g(x) - e$ , where  $e$  has exponential distribution with mean one, and define the slice as  $S = \{x : z < g(x)\}$

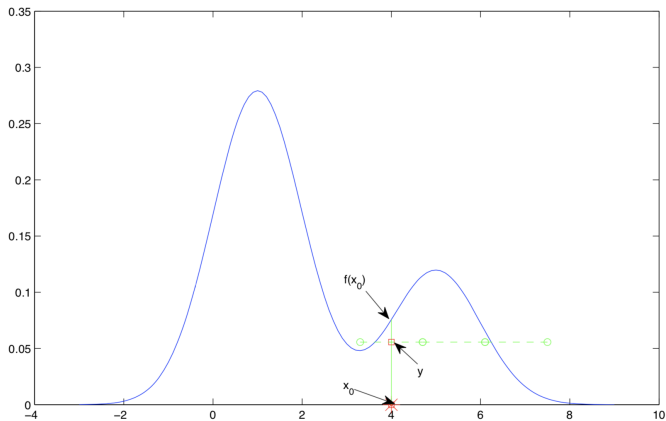
- ▶ One advantage of slice sampling is for sampling from multimodal distributions
- ▶ Unlike standard Metropolis (or MH) that struggles between distant modes, sampling from the slice allows us to easily jump between different modes



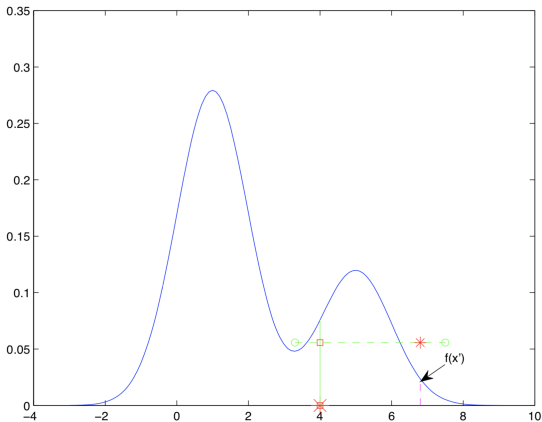


- ▶ Sampling an independent point uniformly from  $S$  might be difficult. In practice, we can substitute this step by any update that leaves the uniform distribution over  $S$  invariant
- ▶ There are several methods to perform this task
- ▶ Here, we introduce a simple but effective procedure that consists of two phases:
  - ▶ *Stepping-out*. A procedure for finding an interval around the current point
  - ▶ *Shrinkage*. A procedure for sampling from the interval obtained
- ▶ For a detail description of these methods, see Neal (2003)

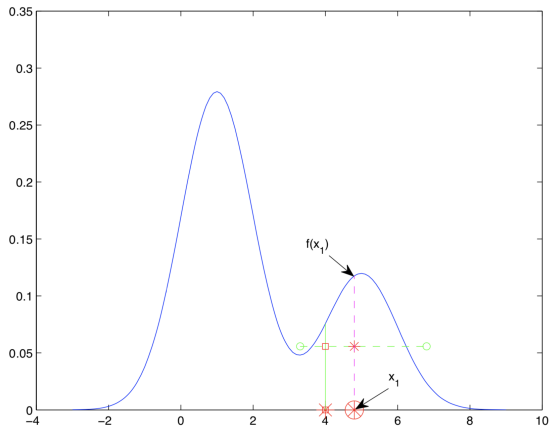
- ▶ Sampling  $u \sim \mathcal{U}(0, f(x_0))$  and stepping out (of size  $w$ ) until we reach points outside the slice



- Shrinkage of interval to a point,  $x'$ , which is sampled (uniformly) from the interval but it has  $f(x') < y$



- ▶ Continue shrinkage until we reach a point  $x_1$  such that  $y < f(x_1)$ . We accept  $x_1$  as our new sample



- ▶ Metropolis, N. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics* 21, 1087–1092.
- ▶ Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57, 97–109.
- ▶ Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6, 721–741.
- ▶ Andrieu, C., De Freitas, N., Doucet, A. and Jordan, M. I. (2003). An introduction to MCMC for machine learning. *Machine learning* 50, 5–43.

- ▶ Griffiths, Thomas L., and Mark Steyvers. 2004. “Finding Scientific Topics.” Proceedings of the National Academy of Sciences of the United States of America 101 (S1): 5228–35.
- ▶ C. J. Geyer (1991) Markov chain Monte Carlo maximum likelihood, Computing Science and Statistics, 23: 156-163.
- ▶ David J. Earl and Michael W. Deem (2005) ”Parallel tempering: Theory, applications, and new perspectives”, Phys. Chem. Chem. Phys., 7, 3910
- ▶ Neal, R. M. Slice sampling. Annals of Statistics, pp. 705–741, 2003.