ECE/ML/CS/ISYE 8803

Approximate Inference in Graphical Models

Module 7: Part C
Metropolis-Hastings Sampling

Faramarz Fekri

Center for Signal and Information Processing

Sampling Based Approximate Inference: So far

- Full particle methods
    - Forward sampling
    - Rejection sampling
    - Weighted likelihood sampling
    - Importance sampling
    - Markov chain Monte Carlo (MCMC)
        - Gibbs sampling
        - Metropolis-Hasting sampling

Today lecture

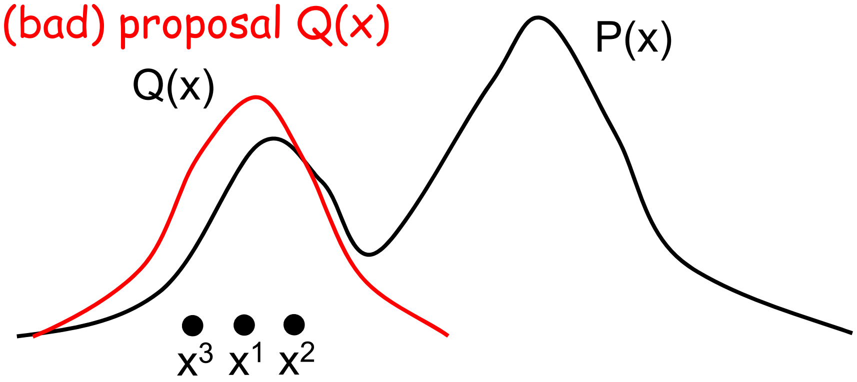
- Distributional particles
    - Rao-Blackwellized particles

Read Chapter 12 of K&F

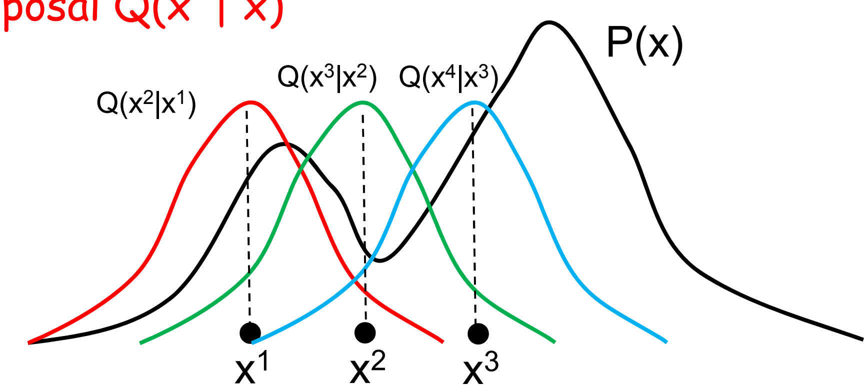
Metropolis-Hastings Sampling

- Instead of a fixed proposal Q(x) (as in Importance Sampling), what if we could use an adaptive proposal?
- Basic idea: We sample from a different distribution Q and then correct for the resulting error.
  - Unlike importance sampling, we do not want to keep track of importance weights as they decay exponentially with number of transitions.
  - Instead, we randomly choose whether to accept the proposed transition, with a probability that corrects for discrepancy between Q and the target distribution P

Importance sampling with a (bad) proposal Q(x)



MCMC with adaptive proposal Q(x' | x)



Metropolis-Hastings Sampling (II)

- Let our proposal distribution T^Q (from which we draw samples) be a transition model over our state space in Markov chain
  - For each state x, T^Q defines a distribution over possible successor states in Val(**X**), from which we select randomly a candidate next state x′

- We can either accept the proposal and transition to the new state x′, or reject it and stay at x.
  - For each states x, for transition to x′, we have an acceptance probability A(x→x′).
  - Actual transition model of the Markov chain is:

$$T(x \to x') = T^Q(x \to x')A(x \to x') \qquad x \neq x'$$
$$T(x \to x) = T^Q(x \to x) + \sum_{x \neq x'} T^Q(x \to x')\big(1 - A(x \to x')\big)$$

Metropolis-Hastings (MH) Sampling (III)

- Using following acceptance probability (and the regularity assumption), the resulting chain T can be shown to have unique stationary $\pi(x) = P(X)$

$$A(x \to x') = \min\left[1, \frac{\pi(x')T^{Q}(x' \to x)}{\pi(x)T^{Q}(x \to x')}\right]$$

- The MH algorithm has a natural implementation in graphical models.
  - Each local transition model T_i is defined via an associated proposal distribution $T_i^{Q_i}$, and the acceptance probability for chain has the form:

$$A(\mathbf{x}_{-i}, x_i \to \mathbf{x}_{-i}, x_i') = \min\left[1, \frac{P(\mathbf{x}_{-i}, x_i')T_i^{Q_i}(\mathbf{x}_{-i}, x_i' \to \mathbf{x}_{-i}, x_i)}{P(\mathbf{x}_{-i}, x_i)T_i^{Q_i}(\mathbf{x}_{-i}, x_i \to \mathbf{x}_{-i}, x_i')}\right]$$

Metropolis-Hastings (MH) Sampling (IV)

- The proposal distributions are usually fairly simple, so it is easy to compute their ratios.
  - In graphical models, the first ratio can also be computed easily:

$$\frac{P(\mathbf{x}_{-i}, x_i')}{P(\mathbf{x}_{-i}, x_i)} = \frac{P(x_i' \mid \mathbf{x}_{-i})}{P(x_i \mid \mathbf{x}_{-i})}$$

Like Gibbs sampling, x_{-i} can be reduced to Markov Blanket of x_{-i}

Metropolis-Hastings (MH) Sampling (V)

- $A(x \rightarrow x')$ is like a ratio of importance sampling weights
  - $P(\mathbf{x}_{-i}, x') / T^Q(x \rightarrow x')$ is the importance weight for x'
  - $P(\mathbf{x}_{-i}, x) / T^Q(x' \rightarrow x)$ is the importance weight for x
  - We divide the importance weight for x' by that of x
- Notice that we only need to compute the ratio rather than $P(\mathbf{x}_{-i}, x')$ or $P(\mathbf{x}_{-i}, x)$ separately.
- Let define $Q(x'|x) = T^Q(x \rightarrow x')$ and $A(x'|x) = A(x \rightarrow x')$
- $A(x'|x) = \min\left(1, \dfrac{P(x')Q(x|x')}{P(x)Q(x'|x)}\right)$ ensures that, after sufficiently many draws, our samples will come from the true distribution P(x) – we shall learn why later.

Metropolis-Hastings Pseudocode

1. Initialize starting state $x^{(0)}$, set $t = 0$
2. Burn-in: while samples have "not converged"
   - $x = x^{(t)}$
   - $t = t + 1$,
   - sample $x^* \sim Q(x^*|x)$ // draw from proposal
   - sample $u \sim \text{Uniform}(0,1)$ // draw acceptance threshold
   - - if $u < A(x^* \mid x) = \min\left(1, \dfrac{P(x^*)Q(x \mid x^*)}{P(x)Q(x^* \mid x)}\right)$
     - $x^{(t)} = x^*$ // transition
   - - else
     - $x^{(t)} = x$ // stay in current state

Function
Draw sample (x(t))

- Take samples from P(x) = : Reset t=0, for $t = 1{:}N$
  - x(t+1) ← Draw sample (x(t))

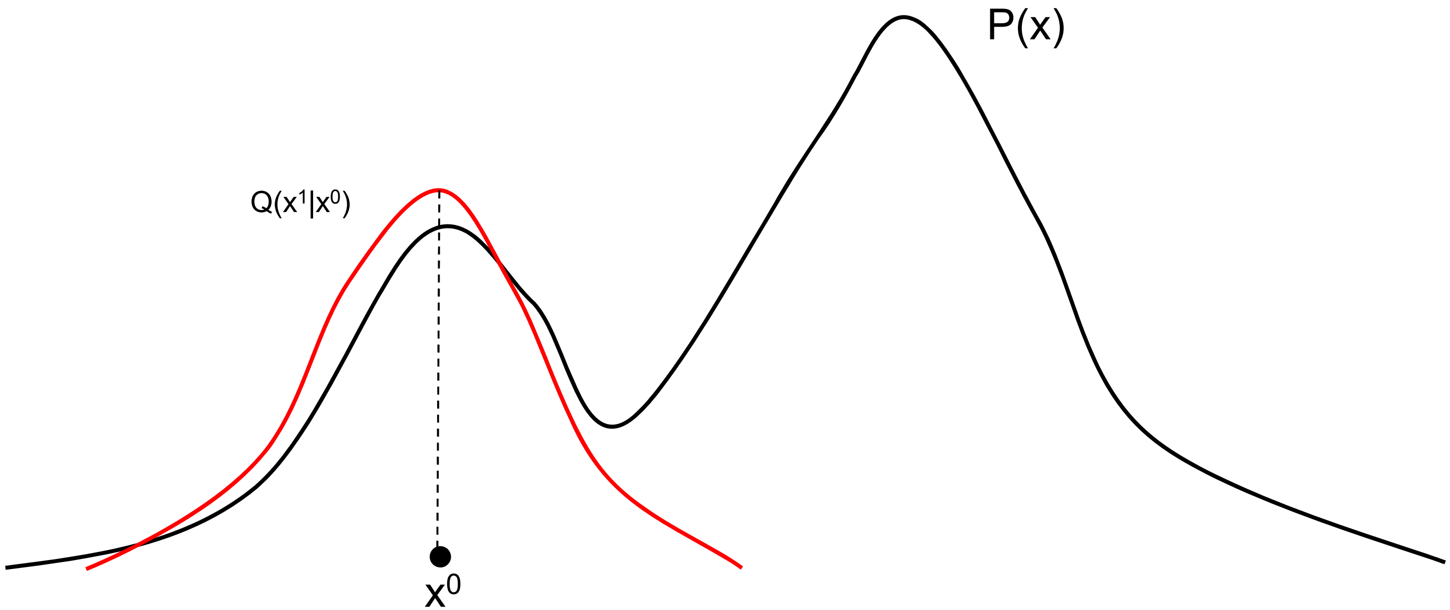
-Xing

Example MH Sampling (I)

- Example:
  - Let Q(x'|x) be a Gaussian centered on x
  - We're trying to sample from a bimodal distribution P(x)

Initialize $x^{(0)}$

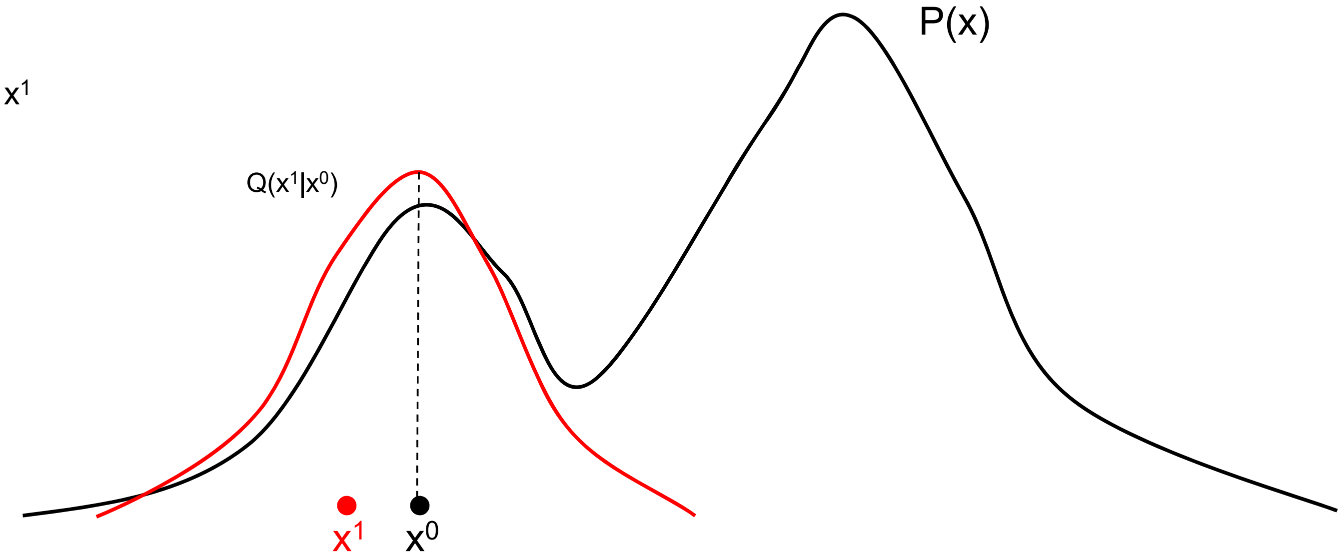
…



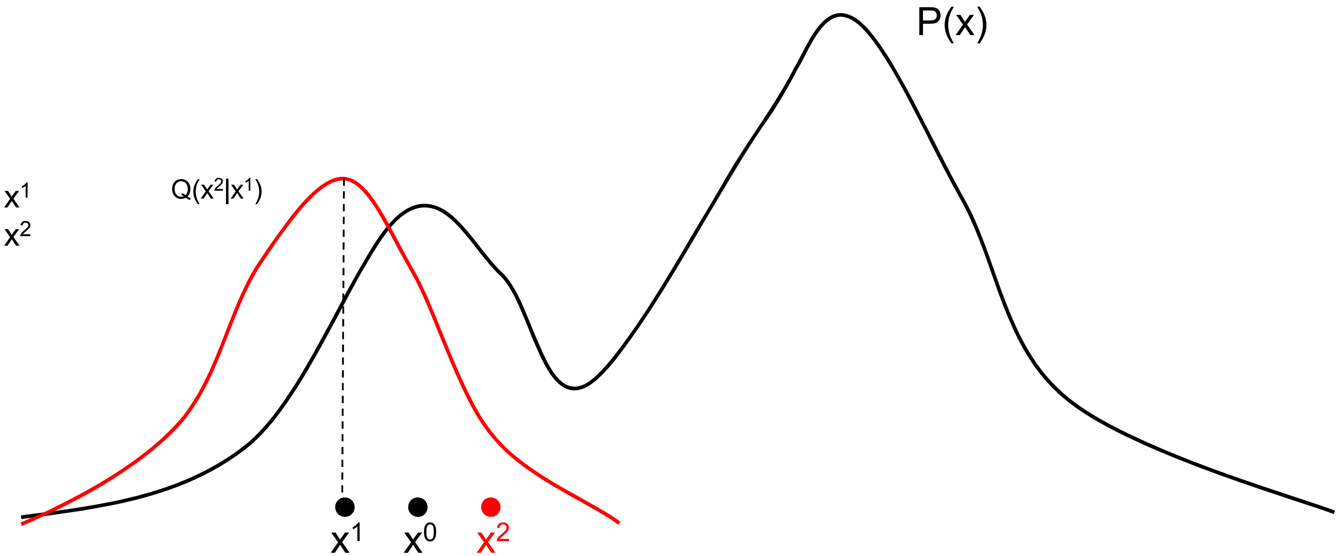
-Xing

Example MH Sampling (II)

Initialize $x^{(0)}$
Draw, accept x^1

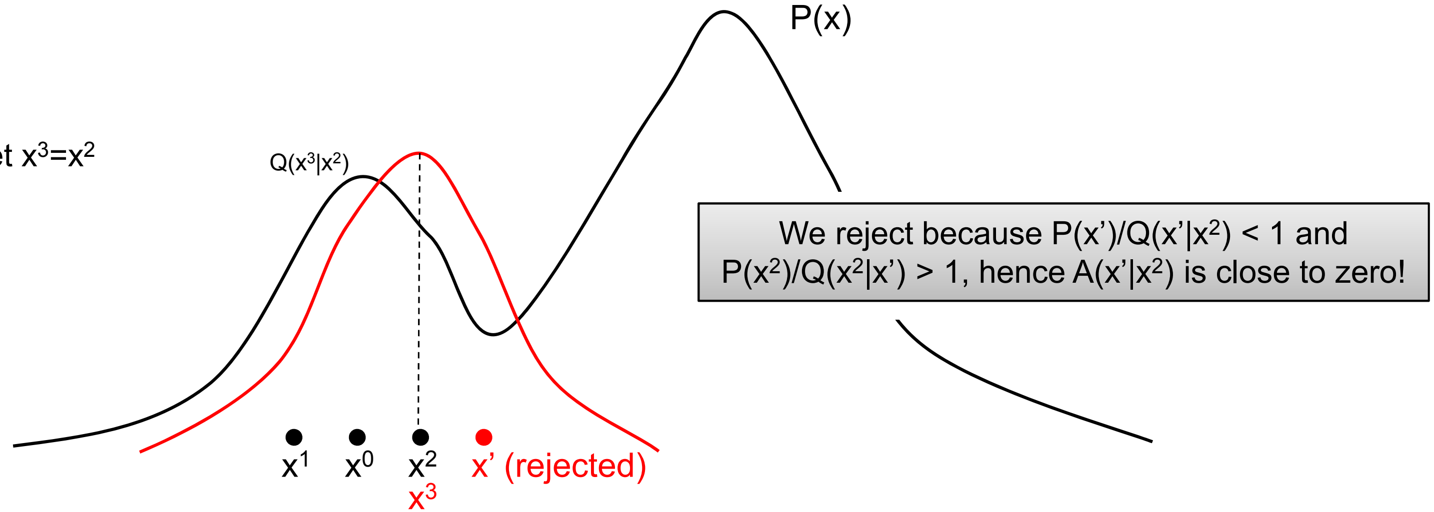


Initialize $x^{(0)}$
Draw, accept x^1
Draw, accept x^2

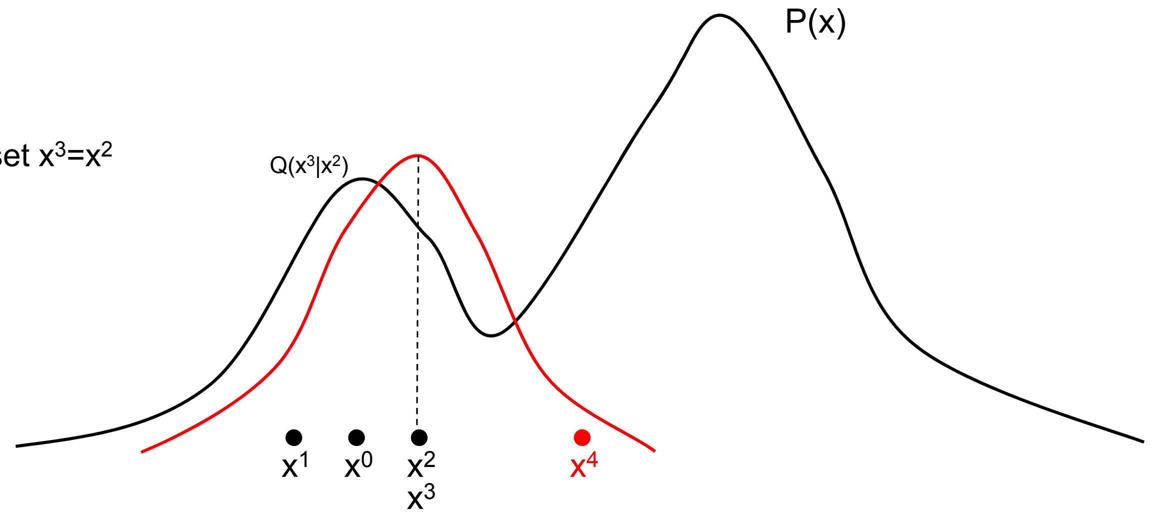


Example MH Sampling (III)

Initialize $x^{(0)}$
Draw, accept x^1
Draw, accept x^2
Draw but reject; set $x^3=x^2$

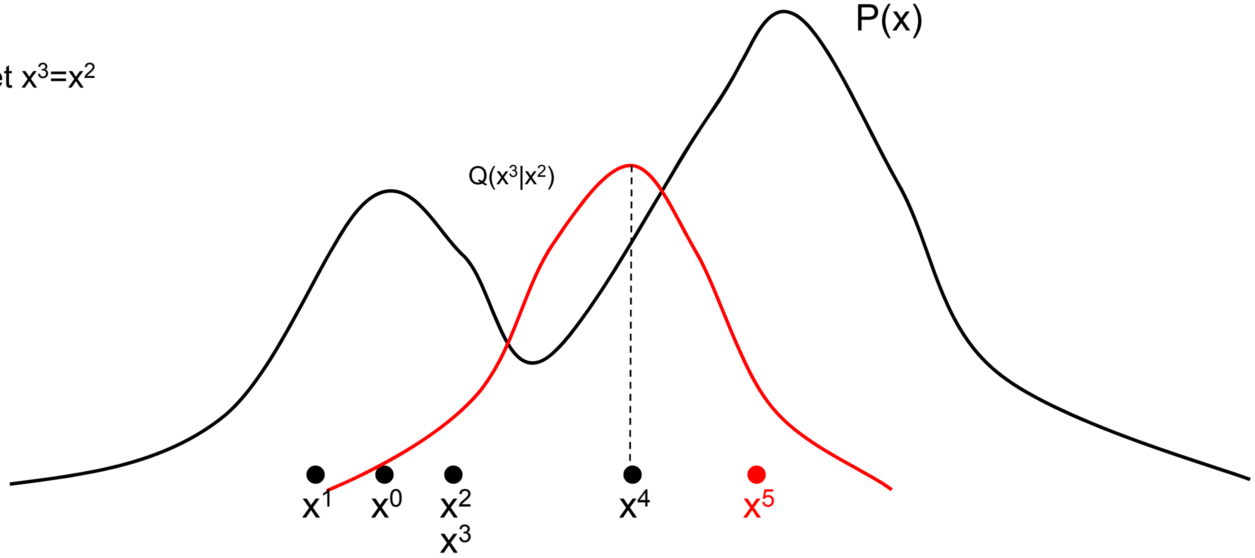


Initialize $x^{(0)}$
Draw, accept x^1
Draw, accept x^2
Draw but reject; set $x^3=x^2$
Draw, accept x^4



Example MH Sampling (IV)

Initialize $x^{(0)}$
Draw, accept x^1
Draw, accept x^2
Draw but reject; set $x^3=x^2$
Draw, accept x^4
Draw, accept x^5



The adaptive proposal Q(x'|x) allows us to sample both modes of P(x)!

Reversible Markov Chain

- Reversible (detailed balance): an MC (with state transition probability $T(x' \to x)$) is reversible if there exists a distribution $\pi(x)$ such that the detailed balance condition is satisfied:

$$\pi(x')T(x' \to x) = \pi(x)T(x \to x')$$

  - Probabilities of x'→x and x→x' are different, but the joint of x and x' is the same, regardless of direction of transition.

Thm: Reversible Markov Chains always have a stationary distribution.

Proof: $$\pi(x')T(x' \to x) = \pi(x)T(x \to x')$$

$$\sum_x \pi(x')\, T(x' \to x) = \sum_x \pi(x)\, T(x \to x')$$

$$\pi(x')\sum_x T(x' \to x) = \sum_x \pi(x)\, T(x \to x')$$

$$\pi(x') = \sum_x \pi(x)\, T(x \to x')$$

- Which is the definition of a stationary distribution in MC.

Reversible Markov Chain (Thm)

Theorem: Reversible Markov Chains always have a stationary distribution.

Proof:

$$\pi(x')T(x' \rightarrow x) = \pi(x)T(x \rightarrow x')$$

$$\sum_x \pi(x')\, T(x' \rightarrow x) = \sum_x \pi(x)\, T(x \rightarrow x')$$

$$\pi(x') \sum_x T(x' \rightarrow x) = \sum_x \pi(x)\, T(x \rightarrow x')$$

$$\pi(x') = \sum_x \pi(x)\, T(x \rightarrow x')$$

- Which is the definition of a stationary distribution in MC.

How does Metropolis-Hastings work?

- Need to prove that MH satisfies detailed balance
  - Note that

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

  - Hence: if $A(x'|x) \le 1$ then $\dfrac{P(x)Q(x'|x)}{P(x')Q(x|x')} \ge 1$ and thus $A(x|x') = 1$

- Suppose A(x'|x) < 1 and A(x|x') = 1. Then we have

$$A(x'|x) = \frac{P(x')Q(x|x')}{P(x)Q(x'|x)}$$

$$P(x)Q(x'|x)A(x'|x) = P(x')Q(x|x')$$

$$P(x)Q(x'|x)A(x'|x) = P(x')Q(x|x')A(x|x')$$

- However, recall: $T(x \to x') = T^Q(x \to x')A(x \to x') = Q(x'|x)A(x'|x)$
- Therefore: $P(x)T(x \to x') = P(x')T(x' \to x)$
- This is the detailed balance condition of MC. Thus P(X) is stationary dist.

Remarks on MH

- P(x) is the true distribution of x.
- MH algorithm converges to a stationary distribution P(x) if

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

- We have no guarantee as to when this convergence to P(x) occurs.
- The burn-in period represents the un-converged part of the Markov Chain – that's why we throw those samples away!
- Like Gibbs sampling, in MH we can resort to mixing time examination, and the plot of the complete log-likelihood vs. time as a way of deciding the convergence.
- MH works better than Gibbs but Gibbs is often the method of choice due to its simplicity.