Approximate inference: Sampling methods

Probabilistic Graphical Models

Tavassolipour

Slides from Dr. Soleymani's PGM course, Sharif University of Technology

Approximate inference

- Approximate inference techniques
  - Deterministic approximation
    - Variational algorithms
  - Stochastic simulation / sampling methods

Sampling-based estimation

- Assume that $\mathcal{D} = \{x^{(1)}, \dots, x^{(N)}\}$ shows the set of i.i.d. samples drawn from the desired distribution P

- For any distribution P, function f, we can estimate $E_P[f]$:

$$E_P[f] \approx \underbrace{\frac{1}{N}\sum_{n=1}^{N} f\left(x^{(n)}\right)}_{\text{Empirical expectation}}$$

- Expectations reveal interesting properties about distribution P
  - Means and variance of P
  - Probability of events
    - E.g., we can find $\hat{P}(x = k)$ by estimating $E_P[f]$ where $f(x) = I(x = k)$

- We can use a stochastic representation of a complex distribution

The mean and variance of the estimator

- For samples drawn independently from the distribution P:

$$\hat{f} = \frac{1}{N}\sum_{n=1}^{N} f\left(x^{(n)}\right)$$

$$E\left[\hat{f}\right] = E[f]$$

$$var\left[\hat{f}\right] = \frac{1}{N}E[(f - E[f])^2]$$

Monte Carlo methods

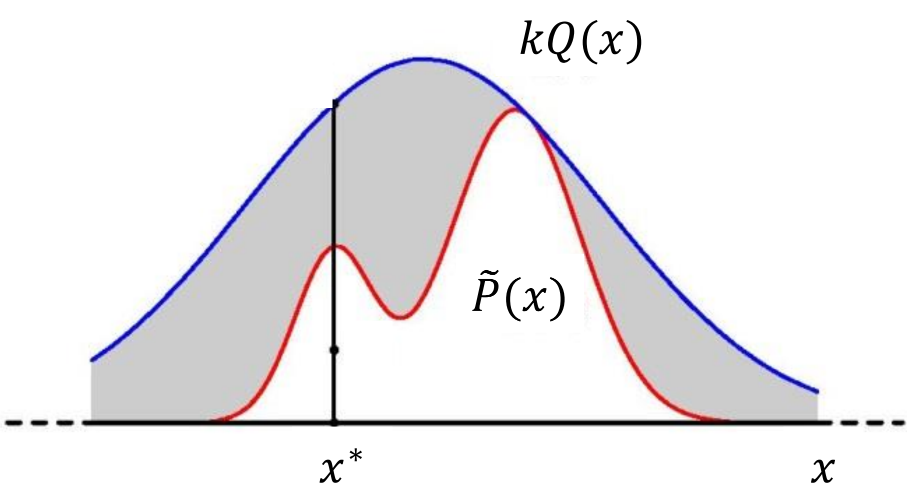
- Using a set of samples to find the answer of an inference query
  - expectations can be approximated using sample-based averages

- **Asymptotically** exact and easy to apply to arbitrary problems

- Challenges:
  - Drawing samples from many distributions is not trivial
  - Are the gathered samples enough?
  - Are all samples useful, or equally useful?

Rejection sampling

- Suppose we wish to sample from $P(\boldsymbol{x}) = \tilde{P}(\boldsymbol{x})/Z$.
  - $P(\boldsymbol{x})$ is difficult to sample, but $\tilde{P}(\boldsymbol{x})$ is easy to evaluate
  - We choose a simpler (proposal) distribution $Q(\boldsymbol{x})$ that we can sample from it more easily
    - Where $\exists k, kQ(\boldsymbol{x}) \geq \tilde{P}(\boldsymbol{x})$
  - Sample from $Q(\boldsymbol{x})$: $\boldsymbol{x}^* \sim Q(\boldsymbol{x})$
  - accept $\boldsymbol{x}^*$ with probability $\frac{\tilde{P}(\boldsymbol{x}^*)}{kQ(\boldsymbol{x}^*)}$



Rejection sampling

- Correctness:

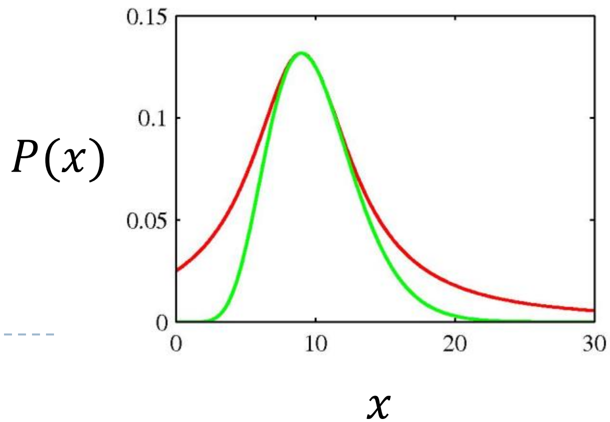
$$\frac{\frac{\tilde{P}(\boldsymbol{x})}{kQ(\boldsymbol{x})}Q(\boldsymbol{x})}{\int \frac{\tilde{P}(\boldsymbol{x})}{kQ(\boldsymbol{x})}Q(\boldsymbol{x})d\boldsymbol{x}} = \frac{\tilde{P}(\boldsymbol{x})}{\int \tilde{P}(\boldsymbol{x})d\boldsymbol{x}} = P(\boldsymbol{x})$$

Probability of acceptance:

$$P(accept) = \int \frac{\tilde{P}(\boldsymbol{x})}{kQ(\boldsymbol{x})}Q(\boldsymbol{x})d\boldsymbol{x} = \frac{\int \tilde{P}(\boldsymbol{x})d\boldsymbol{x}}{k}$$

High dimensional rejection sampling

- Problem: low acceptance rate of rejection sampling in high dimensional spaces
  - exponential decrease of acceptance rate with dimensionality
- Example:
  - Using $Q = N(\boldsymbol{\mu}, \sigma_q^2 \boldsymbol{I})$ to sample $P = N(\boldsymbol{\mu}, \sigma_p^2 \boldsymbol{I})$
  - If σ_q exceeds σ_p by 1%, and $d = 1000$
  - $\left(\frac{\sigma_q}{\sigma_P}\right)^d \approx 20{,}000$ and so the optimal acceptance rate is $1/20{,}000$ that is too small



Limitations of Monte Carlo

- Direct sampling: only when we can sample from $P(\boldsymbol{x})$
  - can be wasteful for rare events

- Rejection sampling uses a proposal distribution $Q(\boldsymbol{x})$ and can also be used when we can not sample $P(\boldsymbol{x})$ directly
  - In rejection sampling, when the proposal $Q(\boldsymbol{x})$ is very different from $P(\boldsymbol{x})$, most samples are rejected

Problem: Finding a good proposal $Q(\boldsymbol{x})$ that is similar to $P(\boldsymbol{x})$ usually requires knowledge of the analytic form of $P(\boldsymbol{x})$ that is not available

Markov Chain Monte Carlo (MCMC)

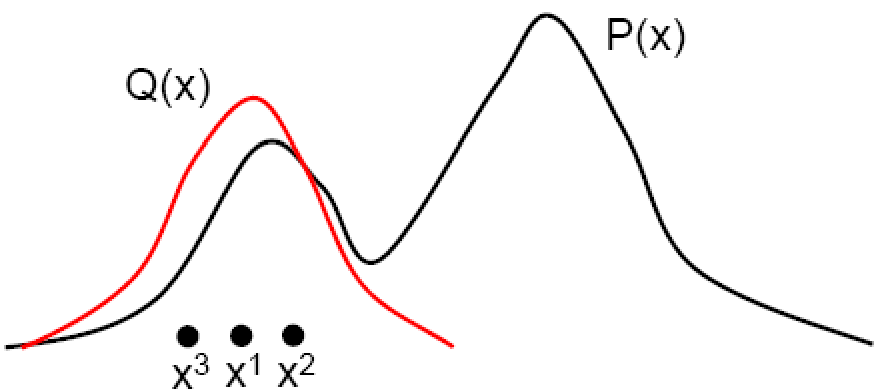
- Instead of using a fixed proposal $Q(\boldsymbol{x})$, we can use an adaptive proposal $Q(\boldsymbol{x}|\boldsymbol{x}^{(t)})$ that depends on the last previous sample $\boldsymbol{x}^{(t)}$
  - The proposal distribution is adapted as a function of the last accepted sample

- MCMC methods
  - Metropolis-Hasting
  - Gibbs

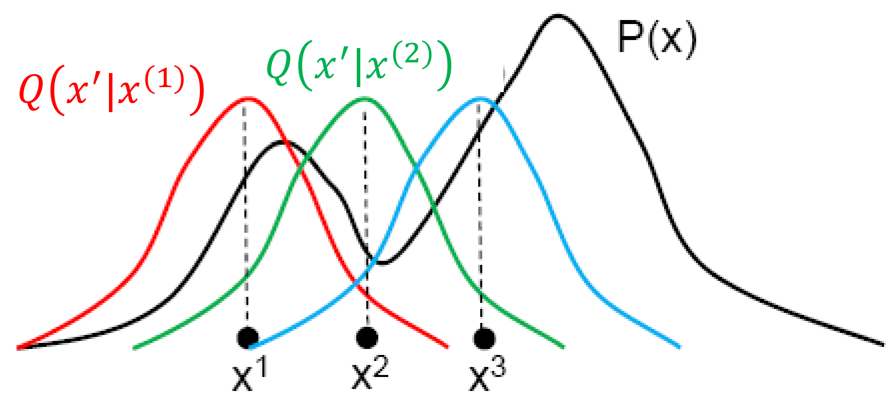
MCMC

- MCMC algorithms feature adaptive proposals
  - Instead of $Q(\boldsymbol{x}')$, they use $Q(\boldsymbol{x}'|\boldsymbol{x})$ where $\boldsymbol{x}'$ is the new state being sampled, and $\boldsymbol{x}$ is the previous sample
- As $\boldsymbol{x}$ changes, $Q(\boldsymbol{x}'|\boldsymbol{x})$ can also change (as a function of $\boldsymbol{x}'$)

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



MCMC with adaptive proposal Q(x'|x)



[This slide has been adopted from Xing's slides]

Markov chains

- A Markov chain is a sequence of random variables $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(t)}$ with the Markov property

$$P\left(\boldsymbol{x}^{(t)} | \boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(t-1)}\right) = P\left(\boldsymbol{x}^{(t)} | \boldsymbol{x}^{(t-1)}\right)$$

- We focus on homogeneous Markov chains, in which $P\left(\boldsymbol{x}^{(t)} | \boldsymbol{x}^{(t-1)}\right)$ is fixed with time
  - Let $\boldsymbol{x}$ be the previous state and $\boldsymbol{x}'$ be the next state, we call $P\left(\boldsymbol{x}^{(t)} | \boldsymbol{x}^{(t-1)}\right)$ as $T(\boldsymbol{x}' | \boldsymbol{x})$
  -
- Thus, at each time point, $\boldsymbol{x}^{(t)}$ is a state showing the configuration of all the variables in the model

Markov chains: invariant or stationary dist.

- $\pi^t(\boldsymbol{x})$: Probability distribution over state $\boldsymbol{x}$, at time t
  - Transition probability $T(\boldsymbol{x}'|\boldsymbol{x})$ redistributes the mass in state $\boldsymbol{x}$ to other states $\boldsymbol{x}'$.

$$\pi^t(\boldsymbol{x}') = \sum_{\boldsymbol{x}} \pi^{t-1}(\boldsymbol{x}) T(\boldsymbol{x}'|\boldsymbol{x})$$

$$T(\boldsymbol{x}'|\boldsymbol{x}) = P\left(\boldsymbol{x}^{(t)} = \boldsymbol{x}' | \boldsymbol{x}^{(t-1)} = \boldsymbol{x}\right)$$

- $\pi(\boldsymbol{x})$ is **invariant** or **stationary** if it does not change under the transitions:

$$\pi(\boldsymbol{x}') = \sum_{\boldsymbol{x}} \pi(\boldsymbol{x}) T(\boldsymbol{x}'|\boldsymbol{x}) \qquad \forall \boldsymbol{x}'$$

Invariant distributions are of great importance in MCMC methods.

More specific than stationary distribution

- There is also no guarantees that the stationary distribution is unique

- In some chains, the stationary distribution reached depends on our starting distribution $\pi^0(\boldsymbol{x})$

- We want to restrict our attention to MCs that have a unique stationary distribution, which is reached from any starting distribution $\pi^0(\boldsymbol{x})$.
  - There are various conditions that suffice to guarantee this property.
  - The most commonly used condition is ergodicity.

Detailed balance

- A sufficient (but not necessary) condition for ensuring that $\pi(\boldsymbol{x})$ is **stationary distribution** of an MC is the **detailed balance** condition:

$$\pi(\boldsymbol{x})T(\boldsymbol{x}'|\boldsymbol{x}) = \pi(\boldsymbol{x}')T(\boldsymbol{x}|\boldsymbol{x}')$$

Detailed balance means the sequences $\boldsymbol{x}', \boldsymbol{x}$ and $\boldsymbol{x}, \boldsymbol{x}'$ are equally probable (although the probability of $\boldsymbol{x}' \to \boldsymbol{x}$ and $\boldsymbol{x} \to \boldsymbol{x}'$ can be different)

Reversible Chains

- Theorem: **Detailed balance** implies the stationary distribution

  Proof:

$$\pi(\boldsymbol{x})T(\boldsymbol{x}'|\boldsymbol{x}) = \pi(\boldsymbol{x}')T(\boldsymbol{x}|\boldsymbol{x}')$$

$$\Rightarrow \sum_{\boldsymbol{x}} \pi(\boldsymbol{x})T(\boldsymbol{x}'|\boldsymbol{x}) = \sum_{\boldsymbol{x}} \pi(\boldsymbol{x}')T(\boldsymbol{x}|\boldsymbol{x}')$$

$$\Rightarrow \sum_{\boldsymbol{x}} \pi(\boldsymbol{x})T(\boldsymbol{x}'|\boldsymbol{x}) = \pi(\boldsymbol{x}') \sum_{\boldsymbol{x}} T(\boldsymbol{x}|\boldsymbol{x}') = \pi(\boldsymbol{x}')$$

- Theorem: If detailed balance holds and T is ergodic, then T has a unique stationary distribution

Stationary distribution: summary

- $\pi^t(\boldsymbol{x})$: Probability distribution over state $\boldsymbol{x}$, at time t
  - Transition probability $T(\boldsymbol{x}'|\boldsymbol{x})$ redistributes the mass in state $\boldsymbol{x}$ to other states $\boldsymbol{x}'$.

$$\pi^{t+1}(\boldsymbol{x}') = \sum_{\boldsymbol{x}} \pi^t(\boldsymbol{x}) T(\boldsymbol{x}'|\boldsymbol{x})$$

- $\pi^*(\boldsymbol{x})$ is **invariant** or **stationary** if it does not change under the transitions:

$$\pi^*(\boldsymbol{x}') = \sum_{\boldsymbol{x}} \pi^*(\boldsymbol{x}) T(\boldsymbol{x}'|\boldsymbol{x}) \;\; \forall \boldsymbol{x}'$$

- A sufficient condition for ensuring that $\pi^*(\boldsymbol{x})$ is **stationary distribution** of an MC is the **detailed balance** condition:

$$\pi^*(\boldsymbol{x}) T(\boldsymbol{x}'|\boldsymbol{x}) = \pi^*(\boldsymbol{x}') T(\boldsymbol{x}|\boldsymbol{x}')$$

How to use Markov chains for sampling from $P(\boldsymbol{x})$?

- Our goal is to use Markov chains to sample from a given distribution.
  - We can achieve this if we set up a Markov chain whose unique stationary distribution is P.
  - We design the transition distribution $T(\boldsymbol{x}'|\boldsymbol{x})$ so that the chain has a unique stationary distribution $P(\boldsymbol{x})$ (independent of P)
    - The ergodic condition is a sufficient condition

Sample $\boldsymbol{x}^{(0)}$ randomly

For t = 0, 1, 2, …
    Sample $\boldsymbol{x}^{(t+1)}$ from $T(\boldsymbol{x}'|\boldsymbol{x}^{(t)})$

Metropolis-Hastings

- Draws a sample $\boldsymbol{x}'$ from $Q(\boldsymbol{x}'|\boldsymbol{x})$, where $\boldsymbol{x}$ is the previous sample

- The new sample $\boldsymbol{x}'$ is accepted with probability $A(\boldsymbol{x}'|\boldsymbol{x})$:

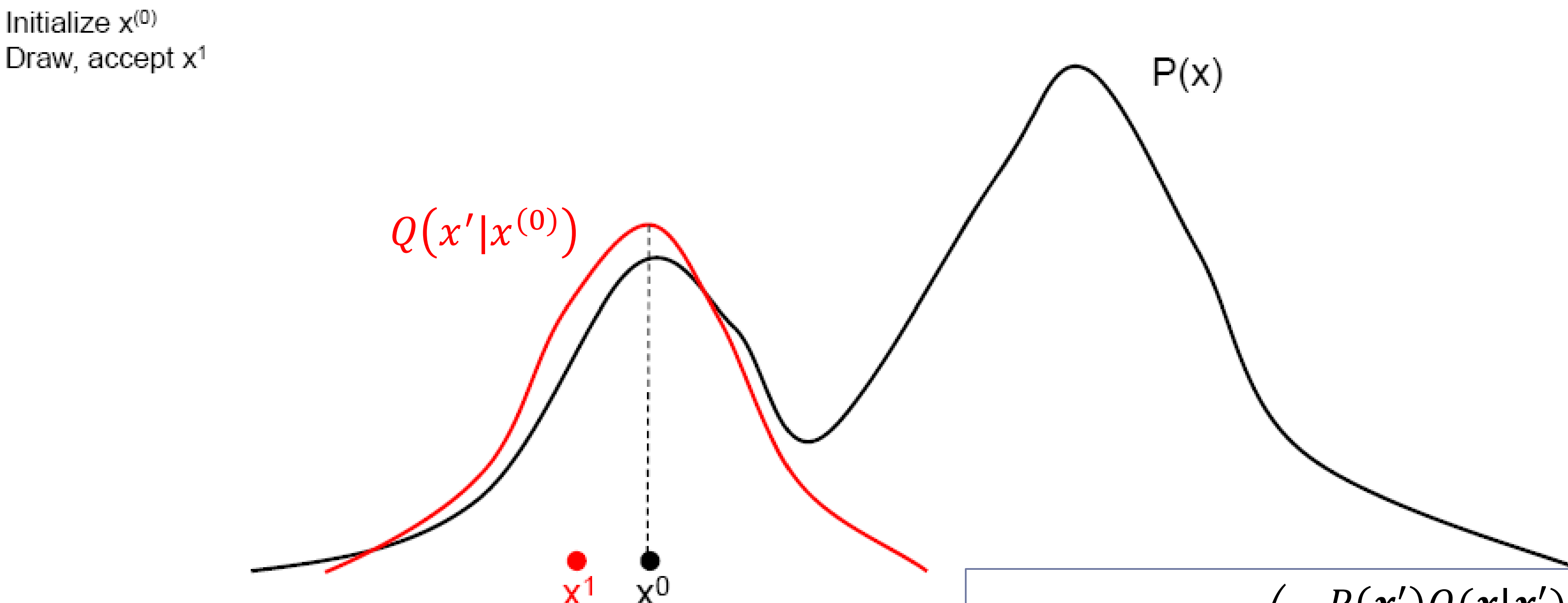
$$A(\boldsymbol{x}'|\boldsymbol{x}) = \min\left(1, \frac{P(\boldsymbol{x}')Q(\boldsymbol{x}|\boldsymbol{x}')}{P(\boldsymbol{x})Q(\boldsymbol{x}'|\boldsymbol{x})}\right)$$

we only need to compute $\frac{P(\boldsymbol{x}')}{P(\boldsymbol{x})}$ rather than $P(\boldsymbol{x}')$ or $P(\boldsymbol{x})$ separately

- We use $A(\boldsymbol{x}'|\boldsymbol{x})$ to ensure that after sufficiently many draws, our samples will come from the true distribution $P(\boldsymbol{x})$

Metropolis-Hastings algorithm: Example

- Let $Q(\boldsymbol{x}'|\boldsymbol{x})$ be a Gaussian centered on $\boldsymbol{x}$
- We're trying to sample from a bimodal distribution $P(\boldsymbol{x})$

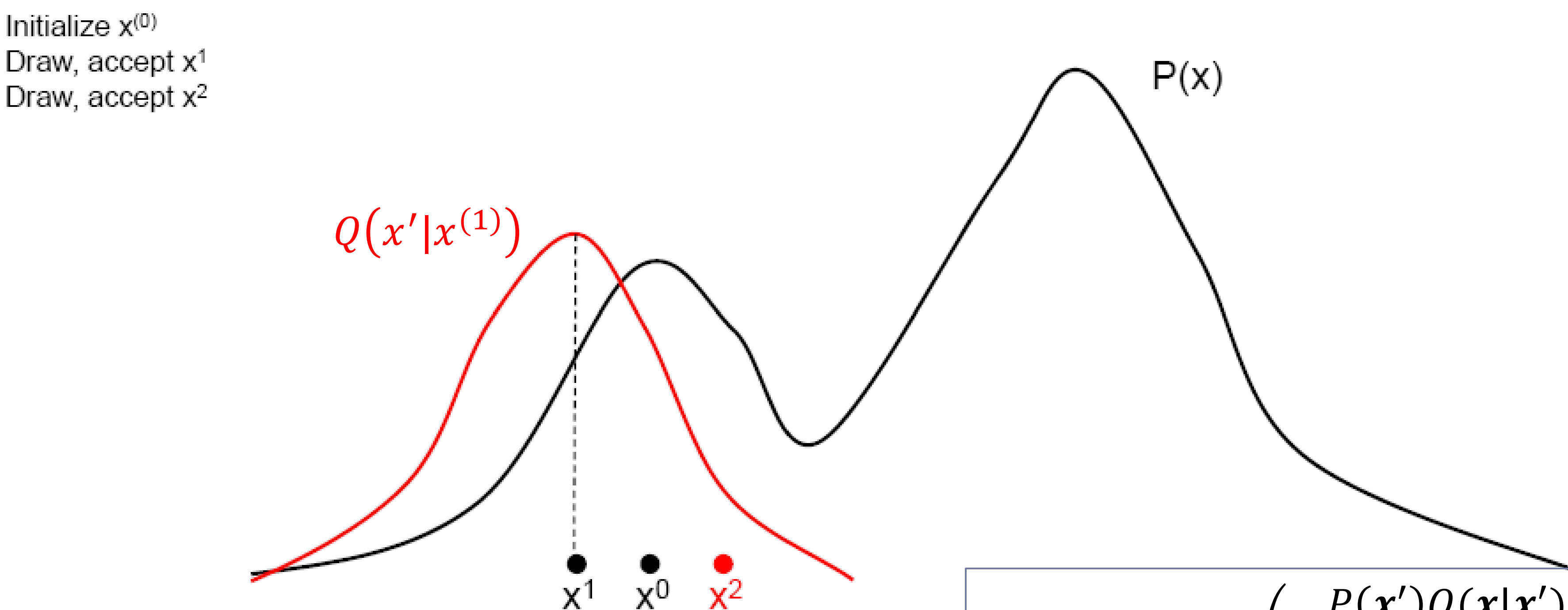


$$A(\boldsymbol{x}'|\boldsymbol{x}) = \min\left(1, \frac{P(\boldsymbol{x}')Q(\boldsymbol{x}|\boldsymbol{x}')}{P(\boldsymbol{x})Q(\boldsymbol{x}'|\boldsymbol{x})}\right)$$

[This slide has been adapted from Xing's slides, CMU]

Metropolis-Hastings algorithm: Example

- Let $Q(\boldsymbol{x}'|\boldsymbol{x})$ be a Gaussian centered on $\boldsymbol{x}$
- We're trying to sample from a bimodal distribution $P(\boldsymbol{x})$



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

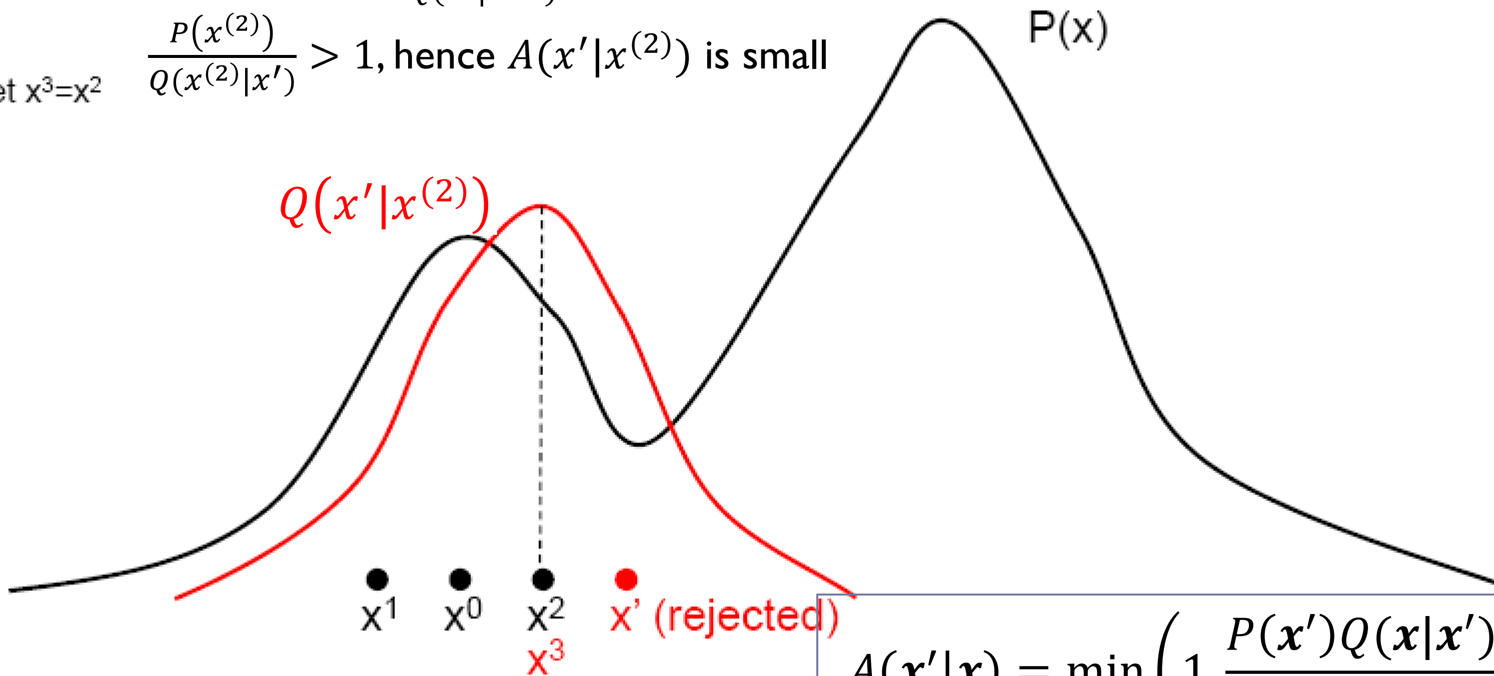
[This slide has been adapted from Xing's slides, CMU]

Metropolis-Hastings algorithm: Example

- Let $Q(\boldsymbol{x}'|\boldsymbol{x})$ be a Gaussian centered on $\boldsymbol{x}$
- We're trying to sample from a bimodal distribution $P(\boldsymbol{x})$

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

Rejected since $\frac{P(x')}{Q(x'|x^{(2)})} < 1$ and $\frac{P(x^{(2)})}{Q(x^{(2)}|x')} > 1$, hence $A(x'|x^{(2)})$ is small

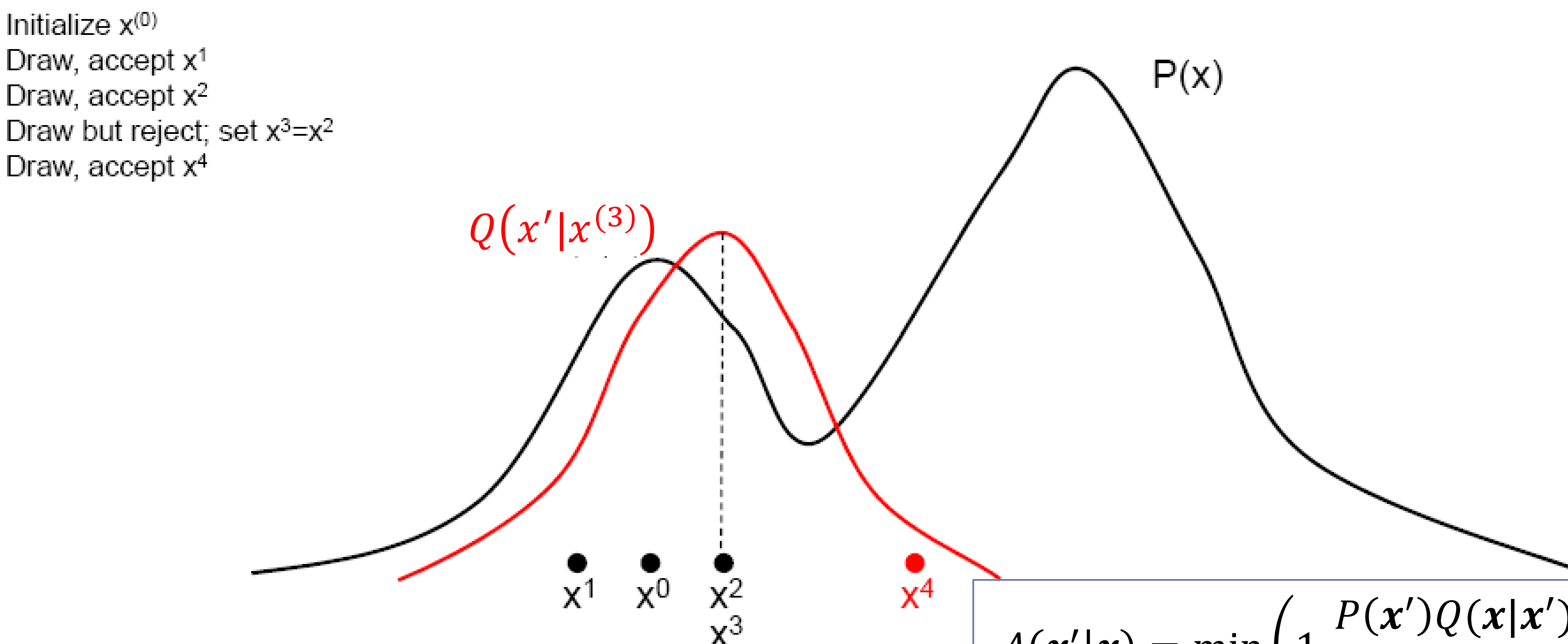


$$A(\boldsymbol{x}'|\boldsymbol{x}) = \min\left(1, \frac{P(\boldsymbol{x}')Q(\boldsymbol{x}|\boldsymbol{x}')}{P(\boldsymbol{x})Q(\boldsymbol{x}'|\boldsymbol{x})}\right)$$

[This slide has been adapted from Xing's slides, CMU]

Metropolis-Hastings algorithm: Example

- Let $Q(\boldsymbol{x}'|\boldsymbol{x})$ be a Gaussian centered on $\boldsymbol{x}$
- We're trying to sample from a bimodal distribution $P(\boldsymbol{x})$



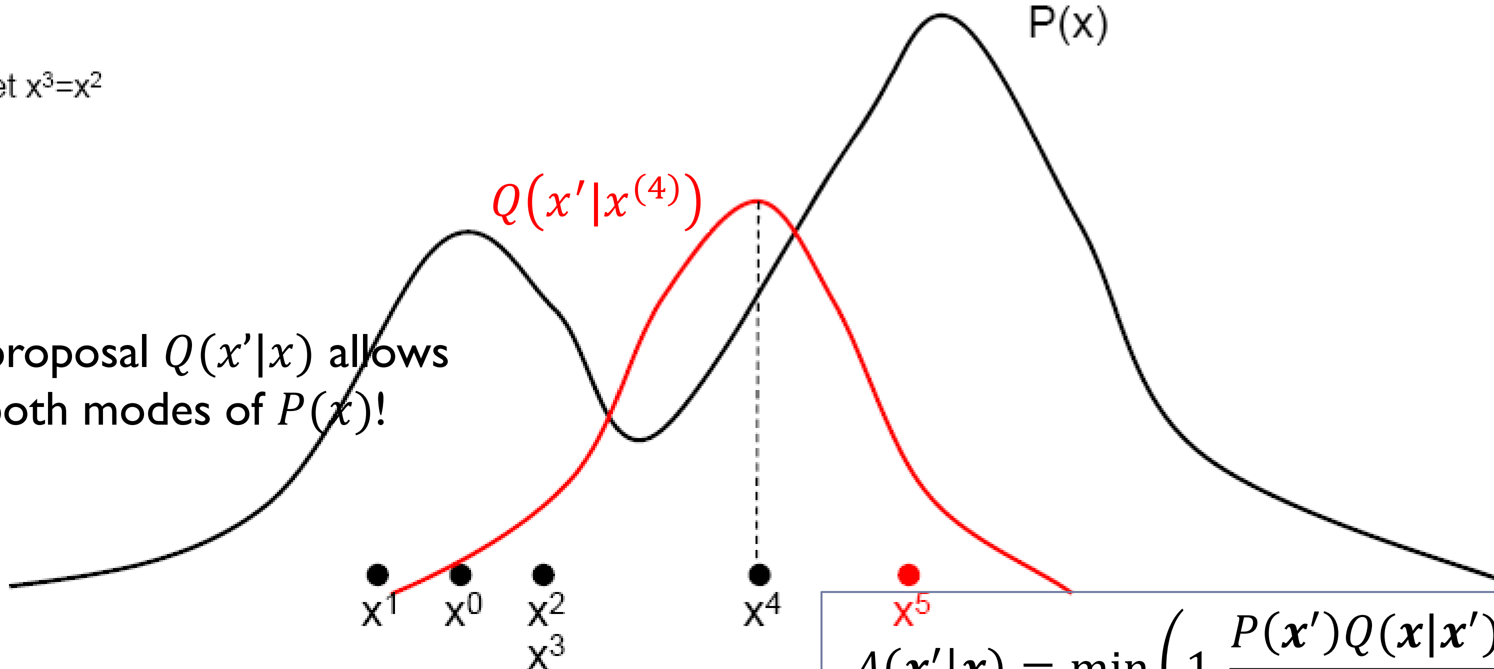
$$A(\boldsymbol{x}'|\boldsymbol{x}) = \min\left(1, \frac{P(\boldsymbol{x}')Q(\boldsymbol{x}|\boldsymbol{x}')}{P(\boldsymbol{x})Q(\boldsymbol{x}'|\boldsymbol{x})}\right)$$

[This slide has been adapted from Xing's slides, CMU]

Metropolis-Hastings algorithm: Example

- Let $Q(\boldsymbol{x}'|\boldsymbol{x})$ be a Gaussian centered on $\boldsymbol{x}$
- We're trying to sample from a bimodal distribution $P(\boldsymbol{x})$

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 $P(x)$!

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

[This slide has been adapted from Xing's slides, CMU]

Metropolis-Hastings algorithm

- Initialize starting state $\boldsymbol{x}^{(0)}$, set $t = 0$
- Burn-in: while samples have "not converged"
  - $\boldsymbol{x} = \boldsymbol{x}^{(t)}$
  - sample $\boldsymbol{x}^* \sim Q(\boldsymbol{x}^*|\boldsymbol{x})$
  - $A(\boldsymbol{x}^*|\boldsymbol{x}) = \min\left(1, \frac{P(\boldsymbol{x}^*)Q(\boldsymbol{x}|\boldsymbol{x}^*)}{P(\boldsymbol{x})Q(\boldsymbol{x}^*|\boldsymbol{x})}\right)$
  - sample $u \sim Uniform(0,1)$
    - if $u < A(\boldsymbol{x}^*|\boldsymbol{x})$
      - $\boldsymbol{x}^{(t+1)} = \boldsymbol{x}^*$
    - else
      - $\boldsymbol{x}^{(t+1)} = \boldsymbol{x}^{(t)}$
  - $t = t + 1$

$\boldsymbol{x}^*$ is accepted with probability $A(\boldsymbol{x}^*|\boldsymbol{x})$

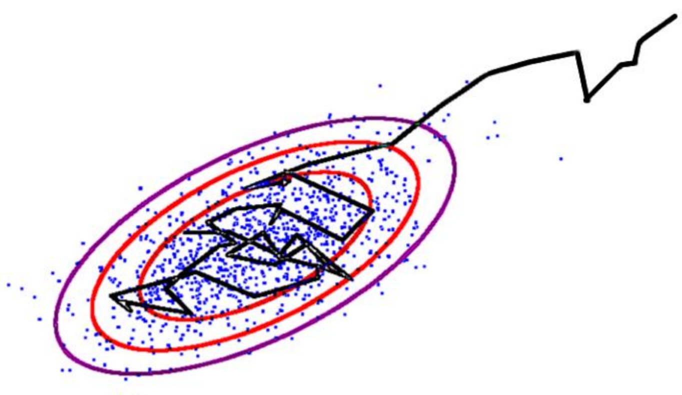
- For n =1…N
  - Draw sample from $Q(\boldsymbol{x}|\boldsymbol{x}^{(n-1)})$ with acceptance probability $A\left(\boldsymbol{x}|\boldsymbol{x}^{(n-1)}\right)$

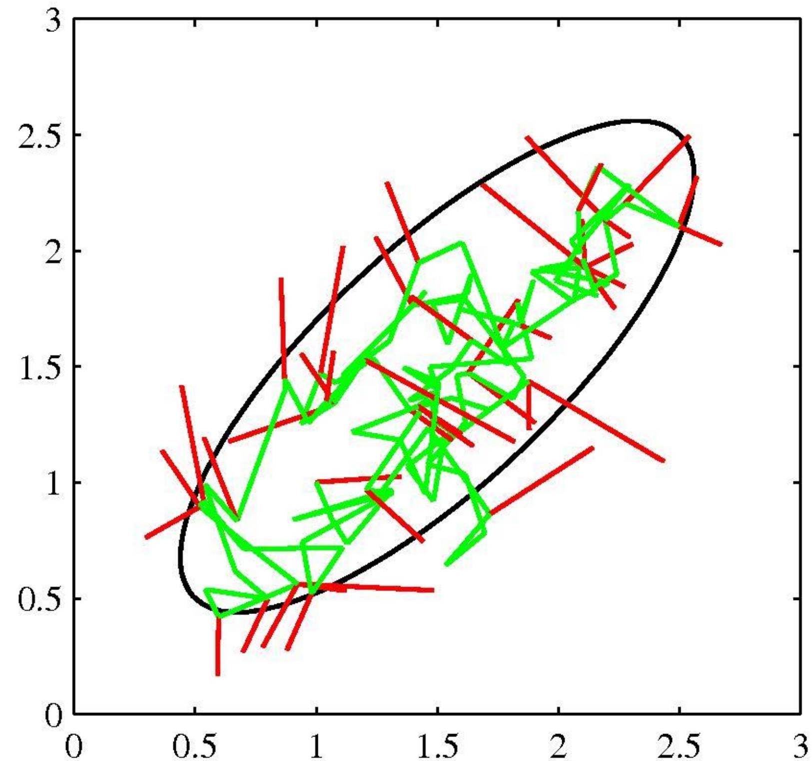
Metropolis algorithm: example

Let $Q(\boldsymbol{x}'|\boldsymbol{x})$ be a Gaussian centered on $\boldsymbol{x}$:

$$Q(\boldsymbol{x}'|\boldsymbol{x}) = N\left(\boldsymbol{x}'\middle|\boldsymbol{x}, \sigma^2\boldsymbol{I}\right)$$

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





A biased random walk that explores the target distribution $P(\boldsymbol{x})$

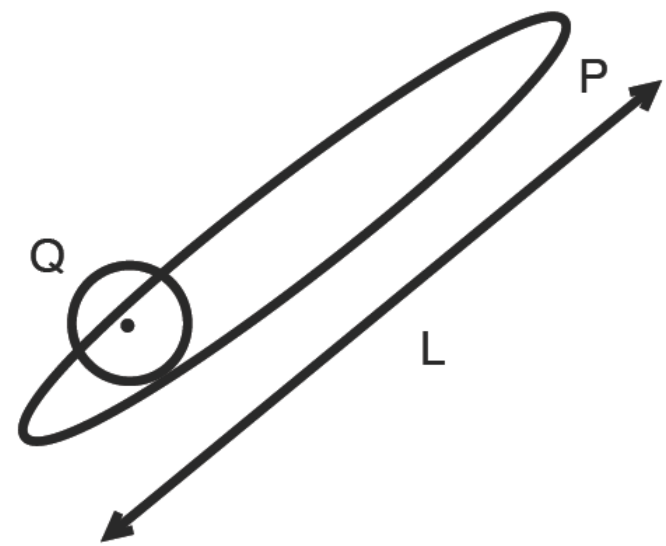
[Bishop book]

Metropolis algorithm: example

- $Q(\boldsymbol{x}'|\boldsymbol{x}) = N\left(\boldsymbol{x}'\middle|\boldsymbol{x}, \sigma^2 \boldsymbol{I}\right)$

- large $\sigma^2 \Rightarrow$ many rejections
- small $\sigma^2 \Rightarrow$ slow exploration
  - $\left(\frac{L}{\sigma}\right)^2$ iterations are required to reach states

In general, finding a good proposal distribution is not always easy



Proposal distribution

- low-variance proposals:
  - high probability of acceptance
  - many iterations are required to explore $P(\boldsymbol{x})$
  - results in more correlated samples

- high-variance proposals
  - low probability of acceptance
  - have the potential to explore much of $P(\boldsymbol{x})$
  - Results in less correlated samples

Theoretical foundation of MH

- Why are the samples generated by MH method will eventually come from $P(\boldsymbol{x})$?
- Why does the MH algorithm have a "burn-in" period?

Why does Metropolis-Hastings work?

- MH is a general construction algorithm that allows us to build a reversible Markov chain with a particular stationary distribution $P(\boldsymbol{x})$

- If we draw a sample $\boldsymbol{x}'$ according to $Q(\boldsymbol{x}'|\boldsymbol{x})$, and then accept/reject according to $A(\boldsymbol{x}'|\boldsymbol{x})$, we have a transition kernel:

$$T(\boldsymbol{x}'|\boldsymbol{x}) = A(\boldsymbol{x}'|\boldsymbol{x})Q(\boldsymbol{x}'|\boldsymbol{x})$$

$$T(\boldsymbol{x}'|\boldsymbol{x}) = A(\boldsymbol{x}'|\boldsymbol{x})Q(\boldsymbol{x}'|\boldsymbol{x}) \quad \text{if } \boldsymbol{x}' \neq \boldsymbol{x}$$

$$T(\boldsymbol{x}|\boldsymbol{x}) = Q(\boldsymbol{x}|\boldsymbol{x}) + \sum_{\boldsymbol{x}' \neq \boldsymbol{x}} Q(\boldsymbol{x}'|\boldsymbol{x})(1 - A(\boldsymbol{x}'|\boldsymbol{x}))$$

MH satisfies detailed balance

- Theorem: MH satisfies detailed balance
- Proof:

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

If $A(\boldsymbol{x}'|\boldsymbol{x}) \leq 1$ then $\frac{P(\boldsymbol{x})Q(\boldsymbol{x}'|\boldsymbol{x})}{P(\boldsymbol{x}')Q(\boldsymbol{x}|\boldsymbol{x}')} \geq 1$ then $A(\boldsymbol{x}|\boldsymbol{x}') = 1$

Suppose that $A(\boldsymbol{x}'|\boldsymbol{x}) < 1$

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

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

$$\xRightarrow{A(\boldsymbol{x}|\boldsymbol{x}')=1} P(\boldsymbol{x})Q(\boldsymbol{x}'|\boldsymbol{x})A(\boldsymbol{x}'|\boldsymbol{x}) = P(\boldsymbol{x}')Q(\boldsymbol{x}|\boldsymbol{x}')A(\boldsymbol{x}|\boldsymbol{x}')$$

$$\Rightarrow P(\boldsymbol{x})T(\boldsymbol{x}'|\boldsymbol{x}) = P(\boldsymbol{x}')T(\boldsymbol{x}|\boldsymbol{x}')$$

$$\boxed{T(\boldsymbol{x}'|\boldsymbol{x}) = A(\boldsymbol{x}'|\boldsymbol{x})Q(\boldsymbol{x}'|\boldsymbol{x})}$$

MH properties

- MH algorithm eventually converges to a stationary distribution $P(\boldsymbol{x})$ that is the true distribution
- However, we have no guarantees as to when this will occur
  - the burn-in period is a way to ignore the un-converged part of the Markov chain
  - but deciding when to halt burn-in is an art that needs experimentation.
- Q must be chosen to fulfill the technical requirements

Gibbs sampling algorithm

Suppose the graphical model contains variables $x_1, \dots, x_M$

Initialize starting values for $x_1, \dots, x_M$

Do until convergence:

Pick an ordering of the M variables (can be fixed or random)

For each variable x_i in order:

Sample x from $P(X_i | \underbrace{x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_M}_{\text{the current values of all other variables}})$

Update $x_i \leftarrow x$

When we update x_i, we immediately use its new value for sampling other variables x_j

Gibbs sampling algorithm

Suppose the graphical model contains variables $x_1, \ldots, x_M$

If the current sample is $\boldsymbol{x} = [x_1, \ldots, x_M]$, the next sample $\boldsymbol{x}' = [x'_1, \ldots, x'_M]$ is drawn as:

Sample x'_1 from $P(X_1 | x_2, \ldots, x_M)$

Sample x'_2 from $P(X_2 | x'_1, x_3, \ldots, x_M)$

$\ldots$

Sample x'_i from $P(X_i | x'_1, \ldots, x'_{i-1}, x_{i+1}, \ldots, x_M)$

$\ldots$

Sample x'_M from $P(X_M | x'_1, \ldots, x'_{M-1})$