# Probabilistic graphical models
# Exact and approximate inference

22-808: Generative models

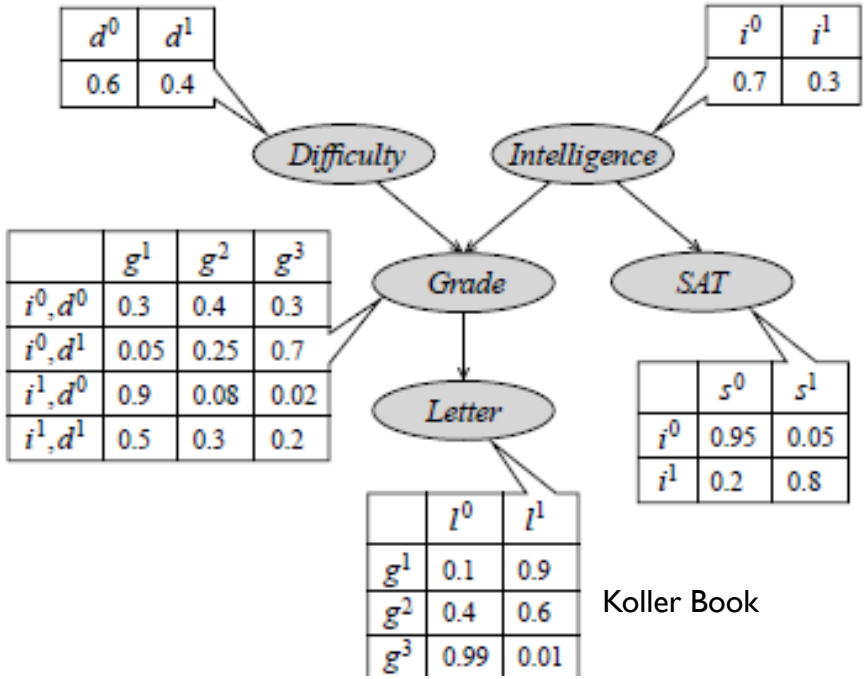
Sharif University of Technology
Fall 2025

Fatemeh Seyyedsalehi

# The inference problem

- Inference: answering **conditional or marginal probabilities in a joint distribution**
  - The graph structure and CPDs (in BNs) or potential functions (in MRFs) are known.
- Example:

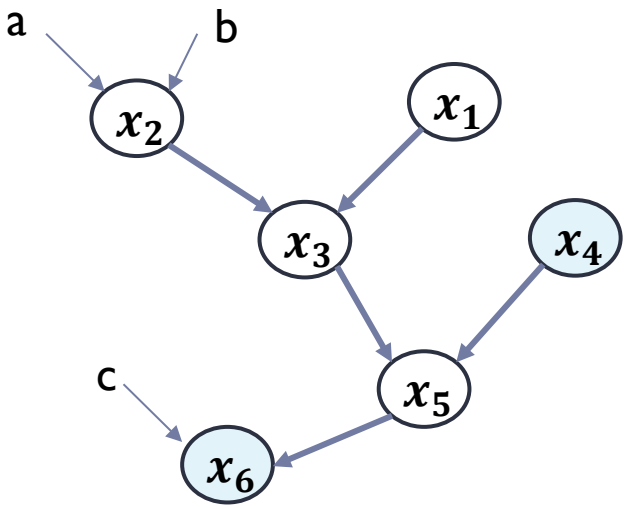
$p(Difficulty|Letter = 1) = ?$
$p(Letter|Intelligence = 1) = ?$
$p(SAT) = ?$




Koller Book

# The inference problem

- Notation:
  - Colored nodes: observed random variables
  - White nodes: latent/hidden/unobserved random variables
  - Others not in a circle: parameters of CPDs

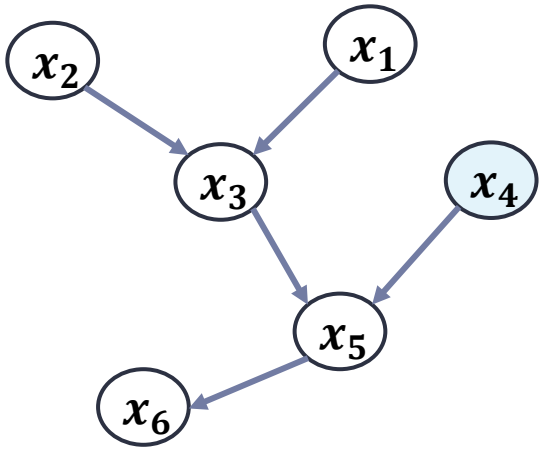


# The inference problem

- Consider the following joint distribution

$$p(x_1, x_2, \dots, x_6)$$

over discrete random variables with *k possible* values.




- An inference query:

$$p(x_2 | x_4 = \overline{x_4}) = ?$$

- A naïve solution:

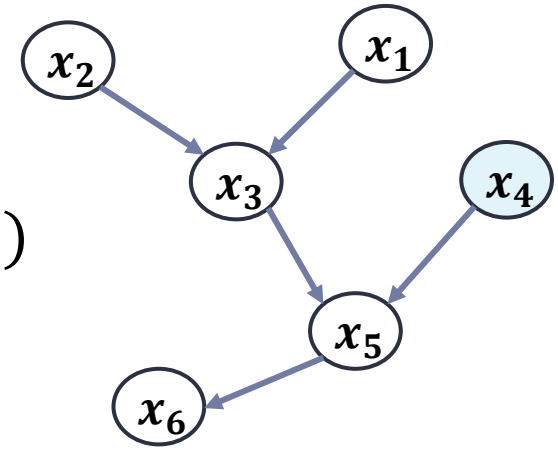
$$p(x_2 | x_4 = \overline{x_4}) = \frac{p(x_2, \overline{x_4})}{\sum_{x_2} p(x_2, \overline{x_4})}$$

$$p(x_2, \overline{x_4}) = \sum_{x_1} \sum_{x_3} \sum_{x_5} \sum_{x_6} p(x_1, x_2, x_3, \overline{x_4}, x_5, x_6)$$

# The inference problem

- However, this distribution is factorized

Over this graph and we have:




$$p(x_1, x_2, \dots, x_6) = p(x_2)p(x_1)p(x_3|x_1, x_2)p(x_5|x_4, x_3)p(x_6|x_5)$$

- A better solution:

$$p(x_2, \overline{x_4}) = \sum_{x_1}\sum_{x_3}\sum_{x_5}\sum_{x_6} p(x_1, x_2, x_3, \overline{x_4}, x_5, x_6)$$

$$= \sum_{x_1}\sum_{x_3}\sum_{x_5}\sum_{x_6} p(x_2)p(x_1)p(x_3|x_1, x_2)p(x_5|\overline{x_4}, x_3)p(x_6|x_5)$$

# The inference problem

- A better solution:

$$p(x_2, \overline{x_4}) = \sum_{x_1}\sum_{x_3}\sum_{x_5}\sum_{x_6} p(x_1, x_2, x_3, \overline{x_4}, x_5, x_6)$$

$$= \sum_{x_1}\sum_{x_3}\sum_{x_5}\sum_{x_6} p(x_2)p(x_1)p(x_3|x_1, x_2)p(x_5|\overline{x_4}, x_3)p(x_6|x_5)$$

- $O(k^4)$ computation !

Distributive law: If $X \notin scope(\phi_1)$ then $\sum_X \phi_1\phi_2 = \phi_1 \sum_X \phi_2$

- Therefore, we can perform summation over the product of only a subset of factors

# The inference problem

- A better solution:

$$p(x_2, \overline{x_4}) = \sum_{x_1}\sum_{x_3}\sum_{x_5}\sum_{x_6} p(x_1, x_2, x_3, \overline{x_4}, x_5, x_6)$$

$$= \sum_{x_1}\sum_{x_3}\sum_{x_5}\sum_{x_6} p(x_2)p(x_1)p(x_3|x_1, x_2)p(x_5|\overline{x_4}, x_3)p(x_6|x_5) =$$

$$p(x_2)\sum_{x_1} p(x_1)\sum_{x_3} p(x_3|x_1, x_2)\sum_{x_5} p(x_5|\overline{x_4}, x_3)\sum_{x_6} p(x_6|x_5)$$

- $O(4k^3)$ computation !

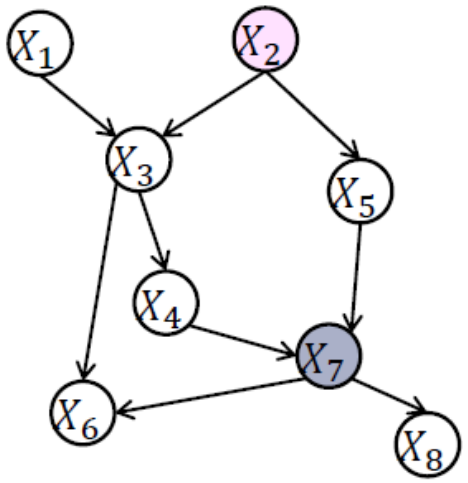
# Variable elimination

- Generally, when a distribution is factorized, Variable elimination algorithm can decrease the computational complexity.

- Variable elimination algorithm for exact inference:
  - We select an elimination order of random variables
  - For each random variable, all factors containing that variable are removed from the set of factors and multiplied
  - The selected random variable is summed out from the product of factors and a new factor is obtained
  - The resulted factor is multiplied to others and algorithm is continued.

# Exact inference: variable elimination Example

- Query: $P(X_2|X_7 = \bar{x}_7)$
- $P(X_2|\bar{x}_7) \propto P(X_2, \bar{x}_7)$




$$P(x_2, \bar{x}_7)$$

$$= \sum_{x_1}\sum_{x_3}\sum_{x_4}\sum_{x_5}\sum_{x_6}\sum_{x_8} P(x_1, x_2, x_3, x_4, x_5, x_6, \bar{x}_7, x_8)$$

Consider the elimination order $X_1, X_3, X_4, X_5, X_6, X_8$

$$P(x_2, \bar{x}_7)$$

$$= \sum_{x_8}\sum_{x_6}\sum_{x_5}\sum_{x_4}\sum_{x_3}\sum_{x_1} P(x_1)P(x_2)P(x_3|x_1, x_2)P(x_4|x_3)P(x_5|x_2)P(x_6|x_3, \bar{x}_7)P(\bar{x}_7|x_4, x_5)P(x_8|\bar{x}_7)$$

# Exact inference: variable elimination Example

$$P(x_2,\bar{x}_7) = \sum_{x_8}\sum_{x_6}\sum_{x_5}\sum_{x_4}\sum_{x_3} P(x_2)P(x_4|x_3)P(x_5|x_2)P(x_6|x_3,\bar{x}_7)P(\bar{x}_7|x_4,x_5)P(x_8|\bar{x}_7)\sum_{x_1} P(x_1)P(x_3|x_1,x_2)$$

$$= \sum_{x_8}\sum_{x_6}\sum_{x_5}\sum_{x_4}\sum_{x_3} P(x_2)P(x_4|x_3)P(x_5|x_2)P(x_6|x_3,\bar{x}_7)P(\bar{x}_7|x_4,x_5)P(x_8|\bar{x}_7)m_1(x_2,x_3)$$

$$= \sum_{x_8}\sum_{x_6}\sum_{x_5}\sum_{x_4} P(x_2)P(x_5|x_2)P(\bar{x}_7|x_4,x_5)P(x_8|\bar{x}_7)\sum_{x_3} P(x_4|x_3)P(x_6|x_3,\bar{x}_7)m_1(x_2,x_3)$$

$$= \sum_{x_8}\sum_{x_6}\sum_{x_5}\sum_{x_4} P(x_2)P(x_5|x_2)P(\bar{x}_7|x_4,x_5)P(x_8|\bar{x}_7)m_3(x_2,x_6,x_4)$$

$$= \sum_{x_8}\sum_{x_6}\sum_{x_5} P(x_2)P(x_5|x_2)P(x_8|\bar{x}_7)\sum_{x_4} P(\bar{x}_7|x_4,x_5)m_3(x_2,x_6,x_4)$$

$$= \sum_{x_8}\sum_{x_6}\sum_{x_5} P(x_2)P(x_5|x_2)P(x_8|\bar{x}_7)m_4(x_2,x_5,x_6)$$

$$= \sum_{x_8}\sum_{x_6} P(x_2)P(x_8|\bar{x}_7)\sum_{x_5} P(x_5|x_2)m_4(x_2,x_5,x_6)$$

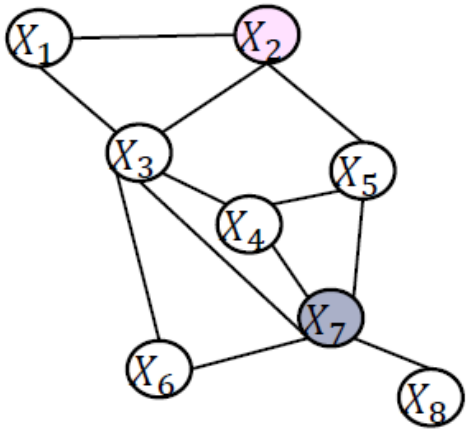
$$= \sum_{x_8}\sum_{x_6} P(x_2)P(x_8|\bar{x}_7)m_5(x_2,x_6)$$

$$= \sum_{x_8} P(x_2)P(x_8|\bar{x}_7)\sum_{x_6} m_5(x_2,x_6)$$

$$= \left(\sum_{x_8} P(x_2)P(x_8|\bar{x}_7)\right)m_6(x_2) = m_8(x_2)m_6(x_2)$$

# Exact inference: variable elimination Example

- Query: $P(X_2|X_7 = \bar{x}_7)$
- $P(X_2|\bar{x}_7) \propto P(X_2, \bar{x}_7)$




$$P(x_2, \bar{x}_7)$$

$$= \sum_{x_1}\sum_{x_3}\sum_{x_4}\sum_{x_5}\sum_{x_6}\sum_{x_8} P(x_1, x_2, x_3, x_4, x_5, x_6, \bar{x}_7, x_8)$$

Consider the elimination order $X_1, X_3, X_4, X_5, X_6, X_8$

$$P(x_2, \bar{x}_7)$$

$$= \sum_{x_8}\sum_{x_6}\sum_{x_5}\sum_{x_4}\sum_{x_3}\sum_{x_1} \phi(x_3, x_4)\phi(x_2, x_5)\phi(x_3, x_6, \bar{x}_7)\phi(x_4, x_5, \bar{x}_7)\phi(\bar{x}_7, x_8)\phi(x_1, x_2, x_3)$$

# Exact inference: variable elimination Example

$$P(x_2,\bar{x}_7) = \sum_{x_8}\sum_{x_6}\sum_{x_5}\sum_{x_4}\sum_{x_3} \phi(x_3,x_4)\phi(x_2,x_5)\phi(x_3,x_6,\bar{x}_7)\phi(x_4,x_5,\bar{x}_7)\phi(\bar{x}_7,x_8)\sum_{x_1}\phi(x_1,x_2,x_3)$$

$$= \sum_{x_8}\sum_{x_6}\sum_{x_5}\sum_{x_4}\sum_{x_3} \phi(x_3,x_4)\phi(x_2,x_5)\phi(x_3,x_6,\bar{x}_7)\phi(x_4,x_5,\bar{x}_7)\phi(\bar{x}_7,x_8)m_1(x_2,x_3)$$

$$= \sum_{x_8}\sum_{x_6}\sum_{x_5}\sum_{x_4} \phi(x_2,x_5)\phi(x_4,x_5,\bar{x}_7)\phi(\bar{x}_7,x_8)\sum_{x_3}\phi(x_3,x_4)\,\phi(x_3,x_6,\bar{x}_7)m_1(x_2,x_3)$$

$$= \sum_{x_8}\sum_{x_6}\sum_{x_5}\sum_{x_4} \phi(x_2,x_5)\phi(x_4,x_5,\bar{x}_7)\phi(\bar{x}_7,x_8)m_3(x_2,x_6,x_4)$$

$$= \sum_{x_8}\sum_{x_6}\sum_{x_5} \phi(x_2,x_5)\phi(\bar{x}_7,x_8)\sum_{x_4}\phi(x_4,x_5,\bar{x}_7)m_3(x_2,x_6,x_4)$$

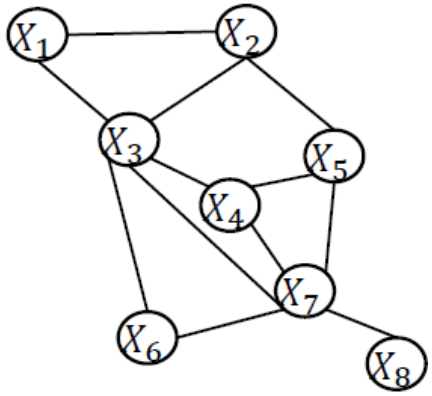
$$= \sum_{x_8}\sum_{x_6}\sum_{x_5} \phi(x_2,x_5)\phi(\bar{x}_7,x_8)m_4(x_2,x_5,x_6)$$

$$= \sum_{x_8}\sum_{x_6} \phi(\bar{x}_7,x_8)\sum_{x_5}\phi(x_2,x_5)m_4(x_2,x_5,x_6)$$

$$= \sum_{x_8}\sum_{x_6} \phi(\bar{x}_7,x_8)m_5(x_2,x_6)$$

$$= \sum_{x_8} \phi(\bar{x}_7,x_8)\sum_{x_6} m_5(x_2,x_6)$$

$$= \left(\sum_{x_8}\phi(\bar{x}_7,x_8)\right)m_6(x_2)$$



# Exact inference: variable elimination

- In each elimination step, we need $O(k^m)$ computations where $m$ is the number of variables is the product of factors containing the variable
  - $k$ is the size of the largest scope of random variables
- With a system with $n$ random variables, we need $O(nk^m)$ computations

# Inference algorithms

- Exact inference :
  - Variable elimination
    - Can be applied on any graph
    - Responds only one query
  - Message passing
    - Sum-product
      - Only for trees
    - Junction tree
      - Can be applied on any graph

- In many real-world applications these algorithms are computationally too complex to be applied

# Inference algorithms

- Approximate inference:
  - Deterministic approximation ←
    - Variational inference
  - Stochastic simulation/sampling methods

# Variational inference

- The calculus of variations (or variational calculus) is a field of mathematical analysis that uses variations, which are small changes in functions and functionals, to find maxima and minima of functionals
  - Functionals: functions of functions ☺

- Variational Bayesian methods are a family of techniques for approximating intractable integrals arising in Bayesian inference and machine learning.

# Variational inference

- Generally, consider two sets of random variables in a joint distribution $p$:
  - $X$: Observed random variables
  - Z: Latent random variables

calculate $p(Z|X)$?

- $p(Z|X) = \frac{p(X,Z)}{p(X)}$, where $p(X) = \int p(X,Z)$
  - We usually have the joint distribution $p(X,Z)$.
  - However, calculating the marginal distribution $p(X)$ is interactable.

# Variational inference

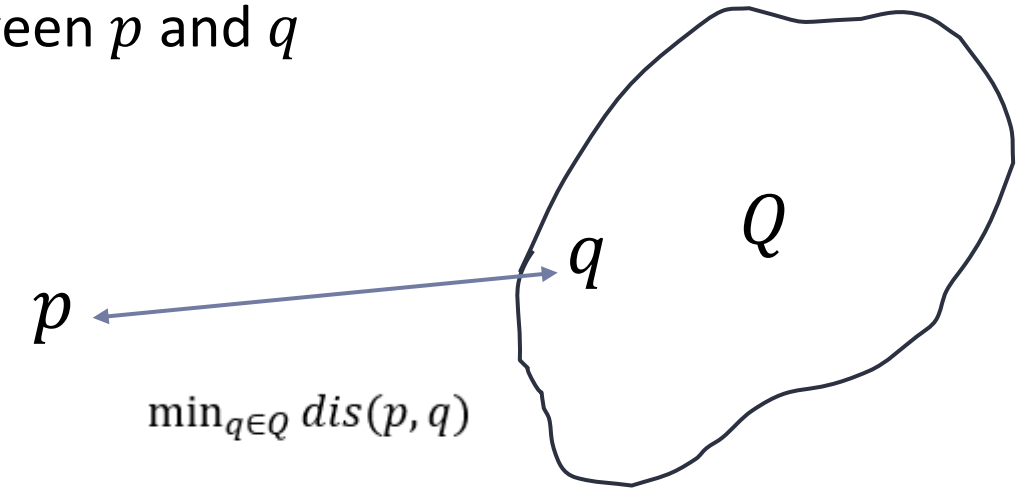
- Generally, consider two sets of random variables in a joint distribution $p$:
  - $X$: Observed random variables
  - Z: Latent random variables

calculate $p(Z|X)$?

- Solution: we select a family distribution $Q$ in which the inference query is tractable. Then we find the best approximate of $p$ in $Q$.

# Variational inference

- Solution: we select a family distribution $Q$ in which the inference query is tractable. Then we find the best approximate of $p$ in $Q$.
- Converting inference to optimization over a functional (variational calculus)
  - A family distribution $Q$
  - A similarity metric between $p$ and $q$
    - KL-divergence



# Variational inference

- Kullback-Leibler divergence between two distribution:

$$KL(p \parallel q) = \int p(x) \log \frac{p(x)}{q(x)} dx$$

- This is positive for any two distributions

- $KL(p||q) = 0$ if and only if $p \equiv q$

- It is not symmetric
  - we call it divergence not distance

# KL-divergence

- Suppose $p$ is the target distribution we want to approximate it,
- I-projection: $KL(q \parallel p)$
- M-projection: $KL(p \parallel q)$
- Obviously, when $p \notin Q$ the result of following optimizations is different

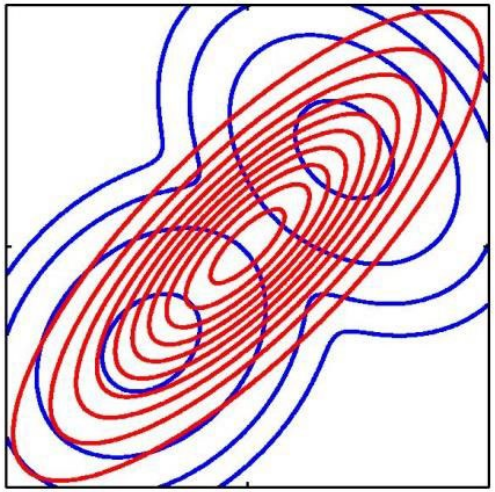
$$\min_q KL(p \parallel q)$$
$$\min_q KL(q \parallel p)$$

# KL-divergence

- $p$ is a mixture of gaussian, $Q$ is the family of gaussian distributions.

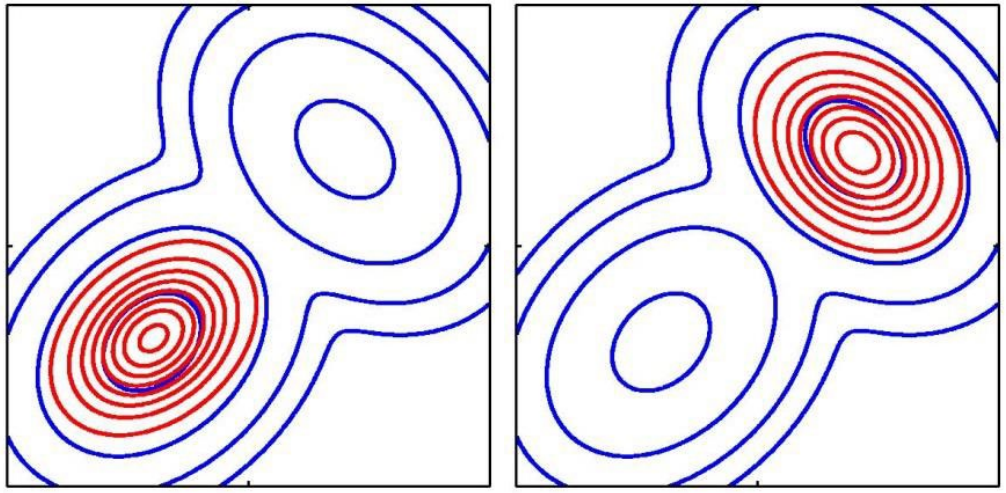
M-projection

$\min_q KL(p \parallel q)$



I-projection

$\min_q KL(q \parallel p)$



Bishop Book

# Variational inference

- In variational inference: I-projection
  - Because of the computational complexity of $p$.

- We should solve the following optimization to find the best $q$:

$$\min_{q} \; KL(q(Z) \parallel p(Z|X))$$

$$KL(q(Z) \parallel p(Z|X)) = \int q(Z) \log \frac{q(Z)}{p(Z|X)} dZ = \int q(Z) \log \frac{q(Z)p(X)}{p(Z,X)} dZ$$

$$= \int q(Z) \log \frac{q(Z)p(X)}{p(Z,X)} dZ = \int q(Z) \log \frac{q(Z)}{p(Z,X)} dZ + \int q(Z) \log p(X) dZ$$

$$= KL\big(q(Z) \parallel p(Z,X)\big) + \log p(x)$$

# Variational inference

- $KL(q(Z) \parallel p(Z|X)) = KL\big(q(Z) \parallel p(Z,X)\big) + \log p(x)$

- Two facts:
  - $KL(q(Z) \parallel p(Z|X)) > 0 \rightarrow \log p(x) > -KL\big(q(Z) \parallel p(Z,X)\big)$

Evidence lower bound (ELBO)

  - $\log p(x) < 0 \rightarrow \ KL\big(q(Z) \parallel p(Z,X)\big) > \ KL(q(Z) \parallel p(Z|X))$

# Variational inference

- $\log p(x) < 0 \rightarrow KL\big(q(Z) \parallel p(Z,X)\big) > KL(q(Z) \parallel p(Z|X))$

The upper bound

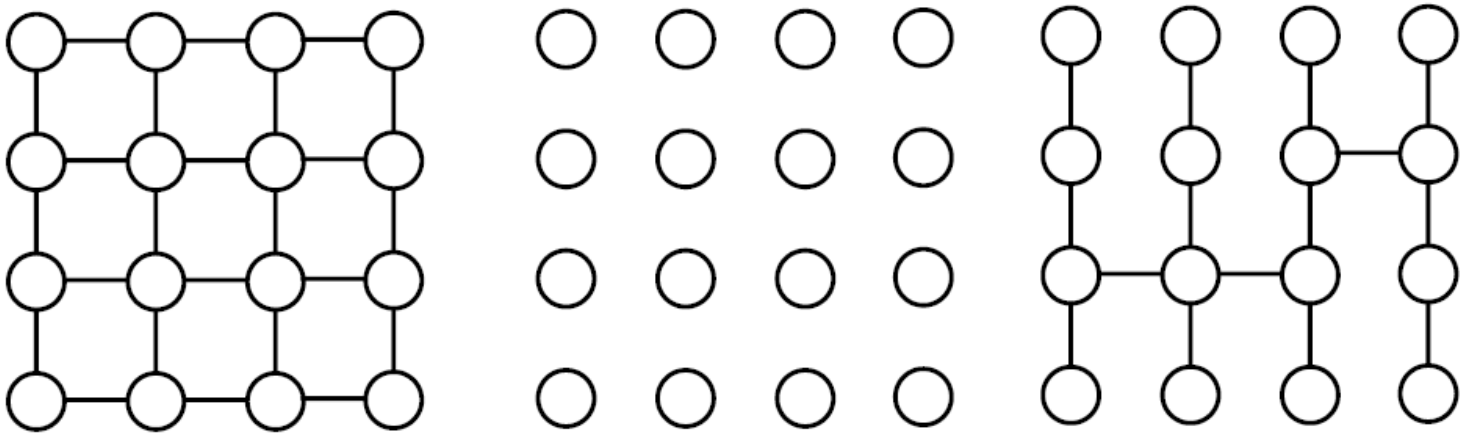
- In variational inference, we minimize the above upper bound

$$\underset{q \in Q}{\operatorname{argmin}}\, KL(q(Z) \parallel p(Z,X))$$

# Variational mean filed approximation

- A common type of variational Bayes

Mean-field assumption: the unknown variables can be partitioned so that each partition is independent of the others



# Variational mean filed approximation

Naïve mean field: the family distribution $Q$ is fully factorized as follows:

$$q(Z) = \prod_i q_i(z_i)$$

- Variational mean field inference:

$$\underset{q \in Q}{\operatorname{argmin}}\, KL(q(Z) \parallel p(Z,X)) = \underset{q_1,q_2,..,q_n}{\operatorname{argmin}}\, KL(q(Z) \parallel p(Z,X))$$

# Variational mean filed approximation

Naïve mean field: the family distribution $Q$ is fully factorized as follows:

$$q(Z) = \prod_j q_j(z_j)$$

- We iteratively optimizing over one coordinate (factor) at a time, as follows,

$$\frac{\partial KL(q(Z)\|p(Z,X))}{\partial q_j} = 0 \text{ to obtain } q_j^*$$

# Variational mean filed approximation

In variational naïve mean field inference, optimum factors are obtained as follows :

$$\log q_j^*(z_j) \propto \mathrm{E}_{q_{-j}}[\log p(X, Z)]$$

As each $q_j^*(z_j)$ depends on others, $q_{-j}$:

1. We initialize $q_j^*$s
2. We iteratively update each $q_j^*$ until convergence!

# Variational mean filed approximation

In variational naïve mean field inference, optimum factors are obtained as follows:

$$\log q_j^*(z_j) \propto \mathrm{E}_{q\_j}[\log p(X,Z)]$$

- Proof:

$$KL\big(q(Z) \parallel p(Z,X)\big) = \int q(Z) \log \frac{q(Z)}{p(Z,X)}\, dZ$$

$$= \mathrm{E}_{q(Z)}\left[\log \frac{q(Z)}{p(X,Z)}\right] = \underbrace{\mathrm{E}_{q(Z)}[\log q(Z)]}_{1} - \underbrace{\mathrm{E}_{q(Z)}[\log p(X,Z)]}_{2}$$

# Variational mean filed approximation proof cont.

- We know,

$$q(Z) = \prod_i q_i(z_i) \rightarrow \log q(Z) = \sum_i \log q_i(z_i)$$

- Therefore,

$$1: \mathrm{E}_{q(Z)}[\log q(Z)] = \sum_i \mathrm{E}_{q_i(z_i)}[\log q_i(z_i)]$$

- Also,

$$2: \mathrm{E}_{q(Z)}[\log p(X,Z)] = \mathrm{E}_{q_j(z_j)}\left[\mathrm{E}_{q_{-j}}[\log p(X,Z)]\right]$$

# Variational mean filed approximation proof cont.

- Therefore,

$$KL\big(q(Z) \parallel p(Z,X)\big)$$
$$= \sum_i \mathrm{E}_{q_i(z_i)}[\log q_i(z_i)] - \mathrm{E}_{q_j(z_j)}\Big[\mathrm{E}_{q_{-j}}[\log p(X,Z)]\Big]$$

- and

$$\frac{\partial\, KL\big(q(Z) \parallel p(Z,X)\big)}{\partial q_j}$$
$$= \frac{\partial \int_{z_j} q_j(z_j)\Big(-\mathrm{E}_{q_{-j}}[\log p(X,Z)] + \log q_j(z_j)\Big)\, dz_j + const}{\partial q_j}$$

# Variational mean filed approximation proof cont.

- According to "Euler–Lagrange equation" we can write,

$$\frac{\partial KL\big(q(Z) \parallel p(Z,X)\big)}{\partial q_j} = \frac{\partial q_j(z_j)\Big(-\mathrm{E}_{q_{-j}}[\log p(X,Z)] + \log q_j(z_j)\Big)}{\partial q_j}$$

- Remember the goal is:

$$q_j^* = \underset{q_j}{\operatorname{argmin}}\, KL\big(q(Z) \parallel p(Z,X)\big)$$

- Therefore:

$$\frac{\partial\, q_j(z_j)\Big(\mathrm{E}_{q_{-j}}[\log p(X,Z)] - \log q_j(z_j)\Big)}{\partial q_j} = 0$$

$$\rightarrow \log q_j^*(z_j) = \mathrm{E}_{q_{-j}}[\log p(X,Z)] + const$$

# Inference algorithms

- Approximate inference:
  - Deterministic approximation
    - Variational inference
  - Stochastic simulation/sampling methods ←

# Sampling based approximation

- Consider the following set of i.i.d. samples from the distribution $p(x)$:

$$D = \{x^1, x^2, \dots, x^n\}$$

Monte Carlo method: for an arbitrary function $f(x)$ of random variable $x$, we can estimate $\mathrm{E}_p[f]$ as follows (empirical expectation),

$$\mathrm{E}_p[f] = \frac{1}{n}\sum_{i=1}^{n} f(x^i)$$

# Sampling based approximation

Monte Carlo method: for an arbitrary function $f(x)$ of random variable $x$, we can estimate $\mathrm{E}_p[f]$ as follows (empirical expectation),

$$\mathrm{E}_p[f] = \frac{1}{n}\sum_{i=1}^{n} f(x^i)$$

- Marginal probability: $\mathrm{p}(\mathrm{x}_1 = \mathrm{k}) \rightarrow f = I(x_1 = k)$
- Mean of a distribution: $f = x$
- …

# Forward sampling in a BN

Given a BN, and number of samples $N$

Choose a **topological** order on variables: $x_1, x_2, \ldots, x_M$

For $i = 1$ to $N$

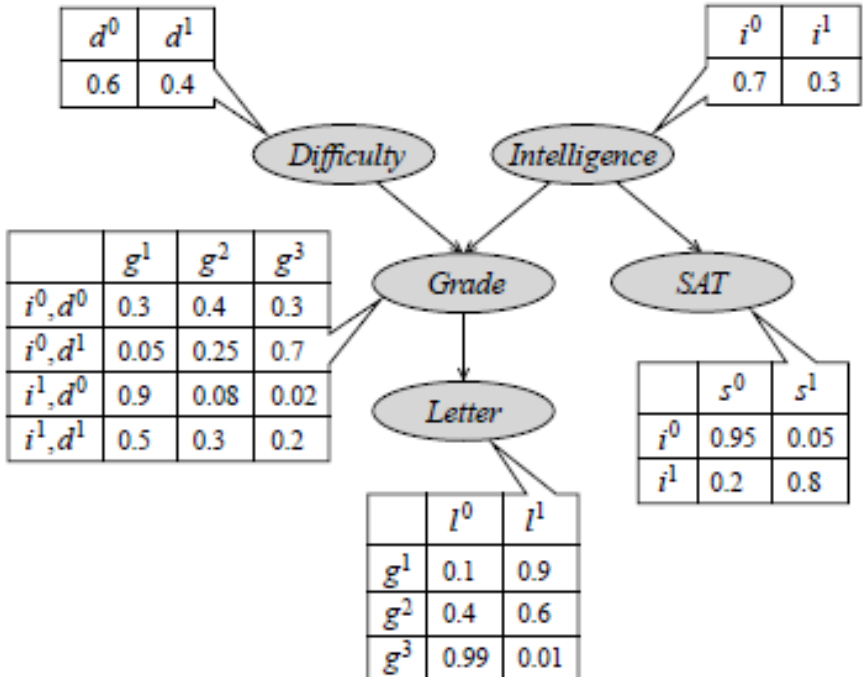
  For $j = 1$ to $M$

    Sample $x_j^i$ from the distribution $p(x_j | parent(x_j))$

  Add $\{x_1^i, x_2^i, \ldots, x_M^i\}$ to the sample set

# Forward sampling in a BN

- $Sample\ D\ from\ p(D)$
- $Sample\ I\ from\ p(I)$
- $Sample\ G\ from\ p(G|D,I)$
- $Sample\ S\ from\ p(S|I)$
- $Sample\ L\ from\ p(L|G)$




Koller Book

# Forward sampling in a BN

- Problems:
  - When the evidence rarely happens, we would need lots of samples, and most would be wasted
  - Overall probability of accepting a sample rapidly decreases when the number of observed variables and states that those variables can take increases
  - This approach is very slow and rarely used in practice.

# Importance sampling

- When sampling from the target distribution $p$ is hard, we use a proposal distribution $q$
  - $q$ should dominates $p \rightarrow q(x) > 0$ whenever $p(x) > 0$

- We sample from the proposal distribution $q$ and consider a weight for each sample:

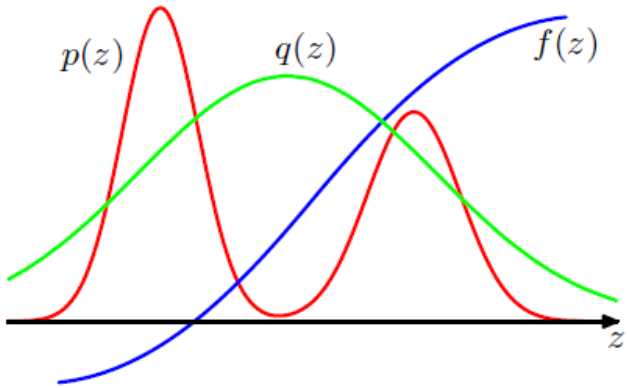
$$\mathrm{E}_p[f] = \int f(x)\, p(x)\, dx = \int f(x) \frac{p(x)}{q(x)} q(x)\, dx$$

$$\mathrm{E}_p[f] \simeq \frac{1}{n} \sum_{i=1}^{n} f\left(x^i\right) \frac{p\left(x^i\right)}{q\left(x^i\right)} \qquad x^i \sim q(x)$$

$$\mathrm{E}_p[f] \simeq \frac{1}{n} \sum_{i=1}^{n} f\left(x^i\right) w\left(x^i\right) \qquad w\left(x^i\right) = \frac{p\left(x^i\right)}{q\left(x^i\right)}$$

# Importance sampling

- Importance sampling depends on how well $q$ matches $p$.
  - For mismatch distributions, weights may be dominated by few samples having large weights, with the remaining weights being relatively insignificant
- It is common that $P(\boldsymbol{x})f(\boldsymbol{x})$ is strongly varying and has a significant proportion of its mass concentrated in a small region
  - The problem is more severe if none of the samples falls in the regions where $P(\boldsymbol{x})f(\boldsymbol{x})$ is large.




Bishop Book

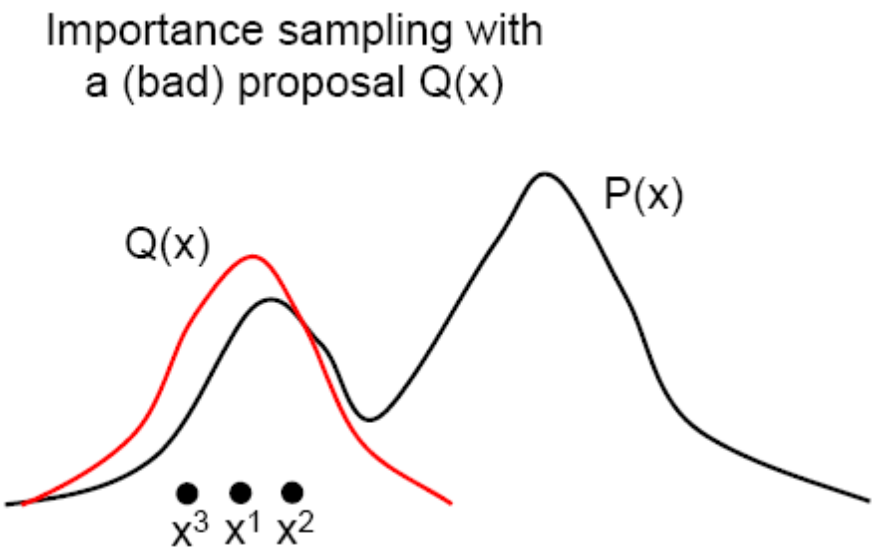
# Problems of naïve Monte Carlo method

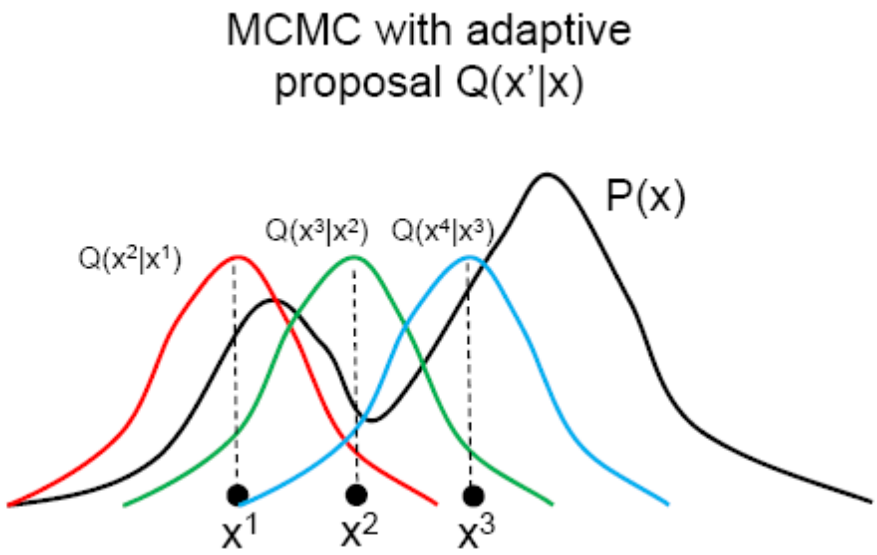
- Direct sampling: only when we can sample from $p(x)$
  - Wasteful for rare evidences

- Importance sampling: when the proposal $q(x)$ is very different from $p(x)$ most samples have very low weights.
  - In fact, finding a good proposal $q(x)$ that is similar to $p(x)$ usually requires knowledge of the analytic form of $p(x)$ that is not available

# Markov chain Monte Carlo (MCMC)

- Using an adaptive distribution $q(x'|x)$ instead of a fixed distribution $q(x)$, where $x$ is the last accepted sample and $x'$ is the new sample.
  - During sampling process the proposal distribution changes as a function of previous sampled data




Bishop Book

# Markov chain Monte Carlo (MCMC)

- Different methods
  - Metropolis-Hastings ←
  - Gibbs sampling

# Markov chain Monte Carlo (MCMC)

- Metropolis-Hastings
  - Sample from $q(x'|x)$, where $x$ is the previous sample
  - As $x$ changes, $q(x'|x)$ can also change
  - Accept this new sample with following probability

$$A(x'|x) = \min(1, \frac{p(x')/q(x'|x)}{p(x)/q(x|x')})$$

importance

  - The acceptance rate $A(x'|x)$ guarantees that after sufficiently many draws, samples are generated from the target distribution $p(x)$
    - Burn-in samples: samples generated in initial iterations and are not from $p(x)$

# Markov chain Monte Carlo (MCMC)

Metropolis-Hastings algorithm:

Initialize starting point: $x^0$ , set t=0

Repeat until convergence:

Sample $x^* \backsim q(x^*|x)$

$$A(x^*|x) = \min(1, \frac{p(x')/q(x'|x)}{p(x)/q(x|x')})$$

Sample $u \backsim uniform\ (0,1)$

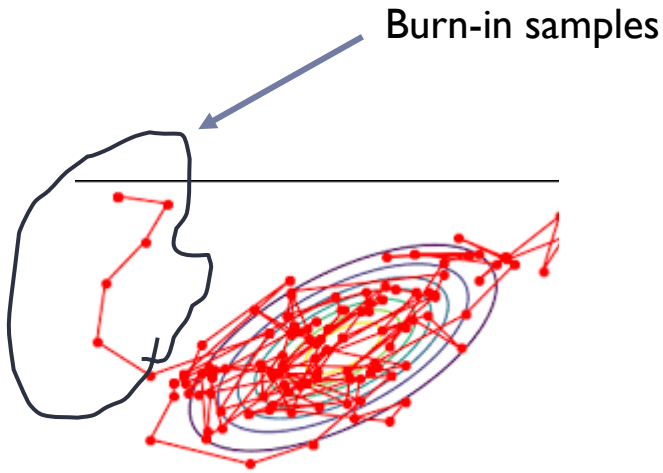
If $u < A(x^*|x)$:

$$x^{t+1} = x^*$$

Else:

$$x^{t+1} = x^t$$

Discard Burn-in samples

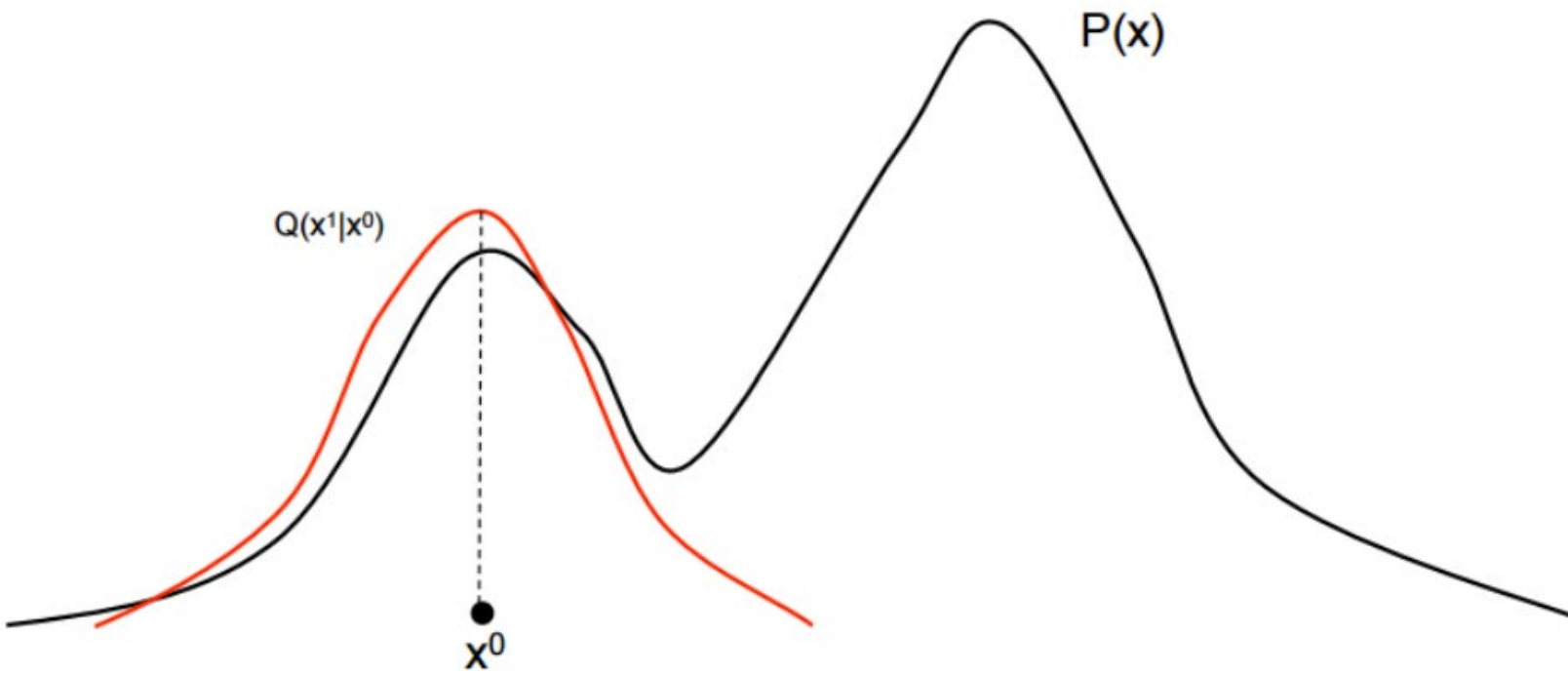


# Markov chain Monte Carlo (MCMC)

- Metropolis-Hastings example:
  - The proposal distribution $q(x'|x)$ is a gaussian distribution
  - The true distribution $p(x)$ is a bimodal with two peaks!

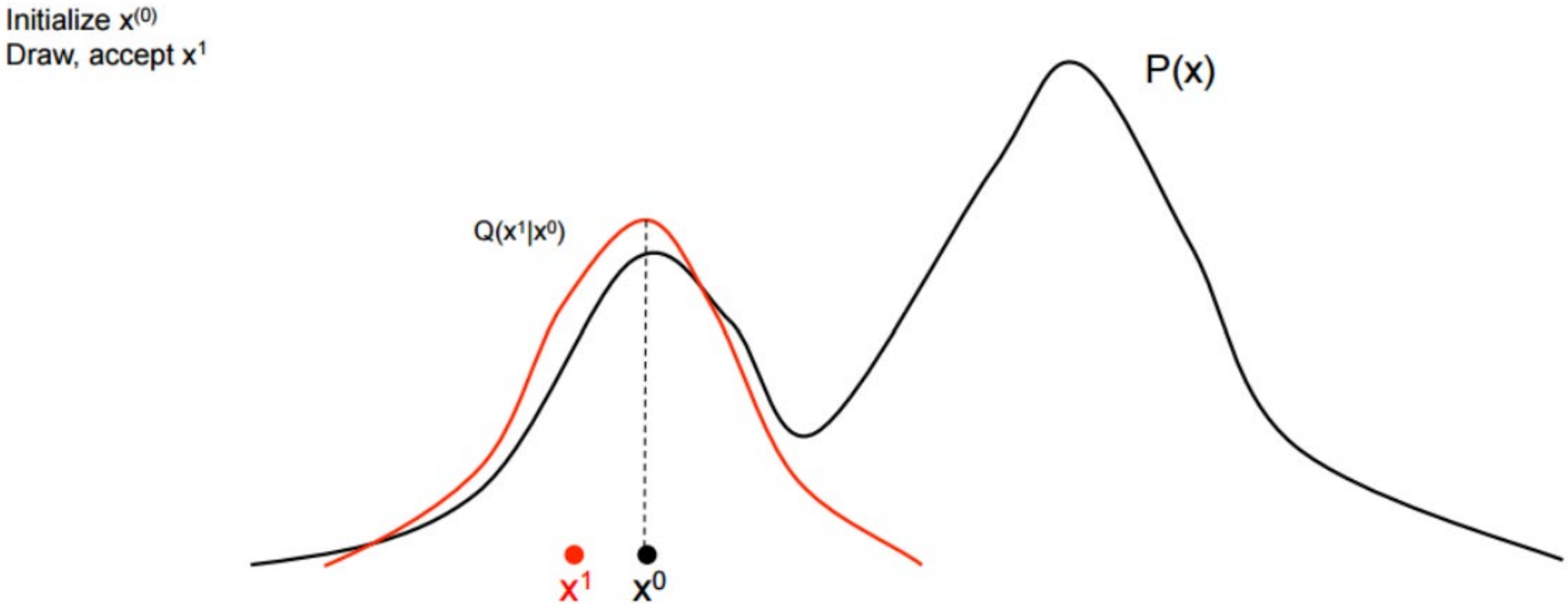
Initialize $x^{(0)}$

…



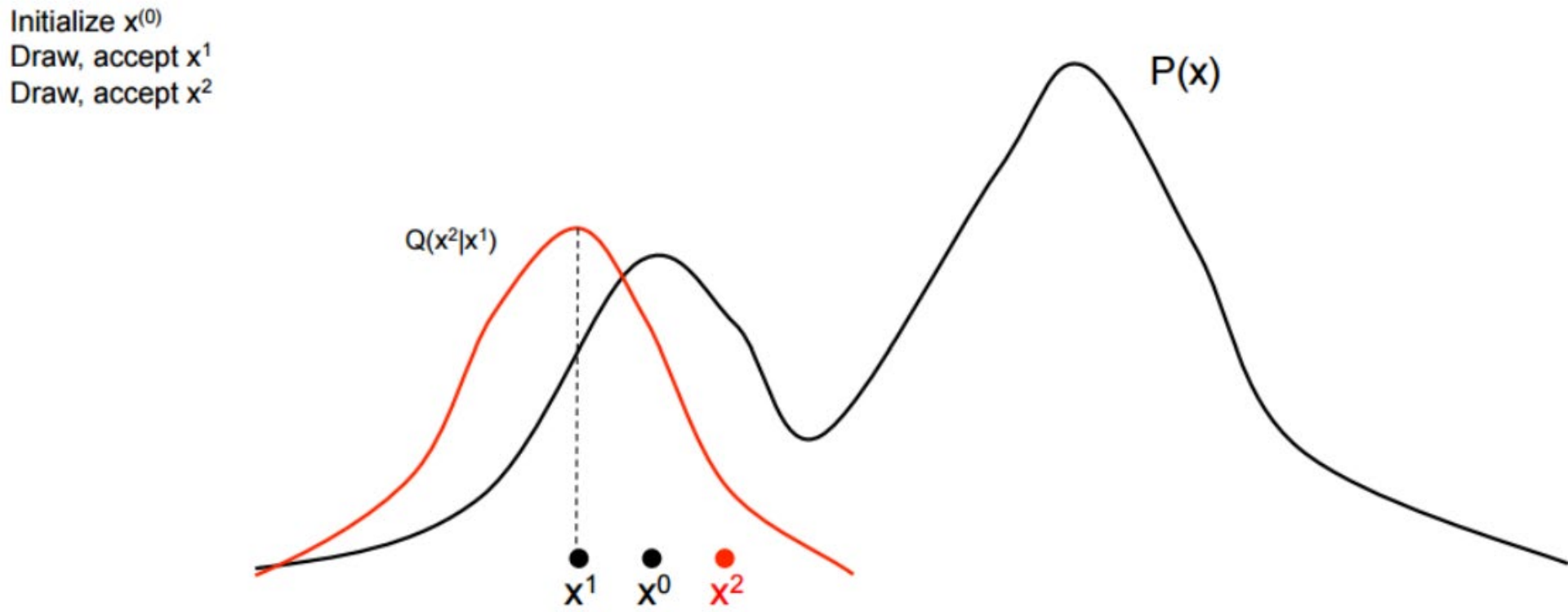
# Markov chain Monte Carlo (MCMC)

- Metropolis-Hastings example:



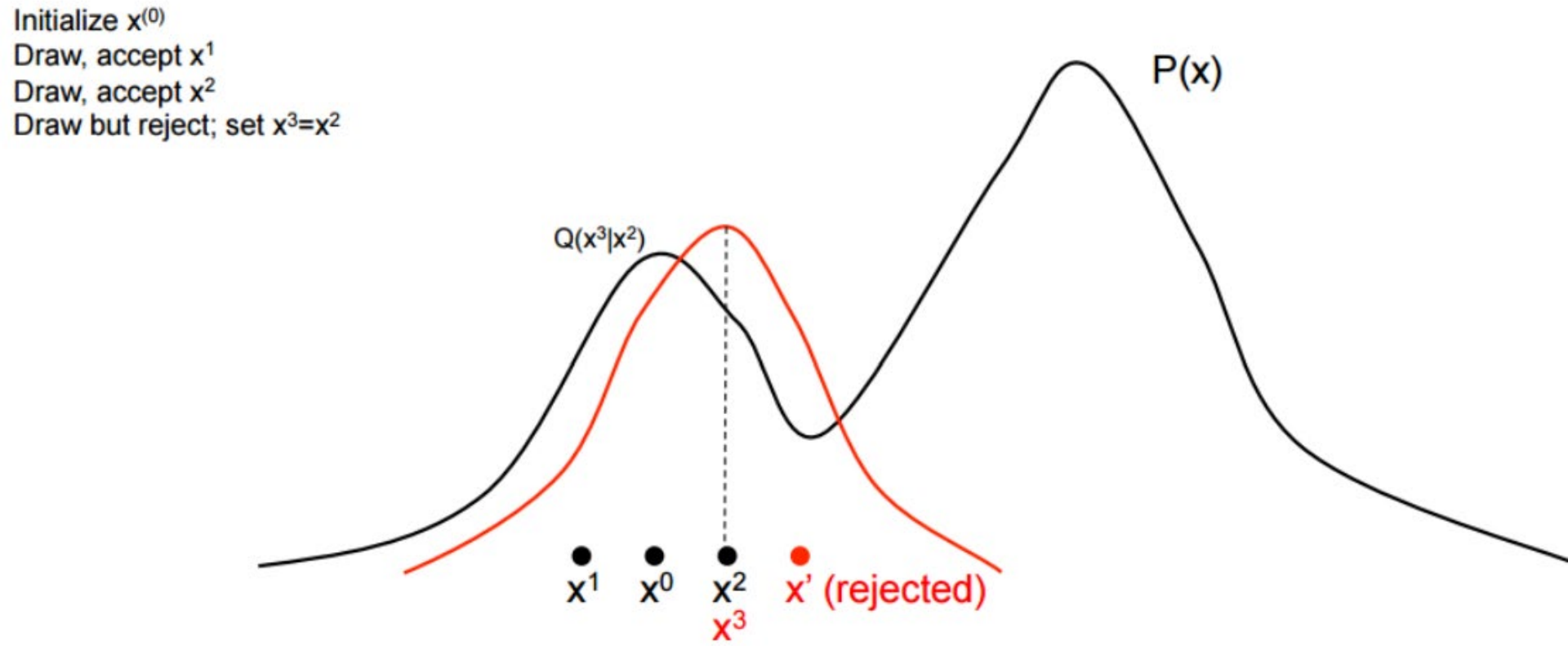
# Markov chain Monte Carlo (MCMC)

- Metropolis-Hastings example:



# Markov chain Monte Carlo (MCMC)

- Metropolis-Hastings example:

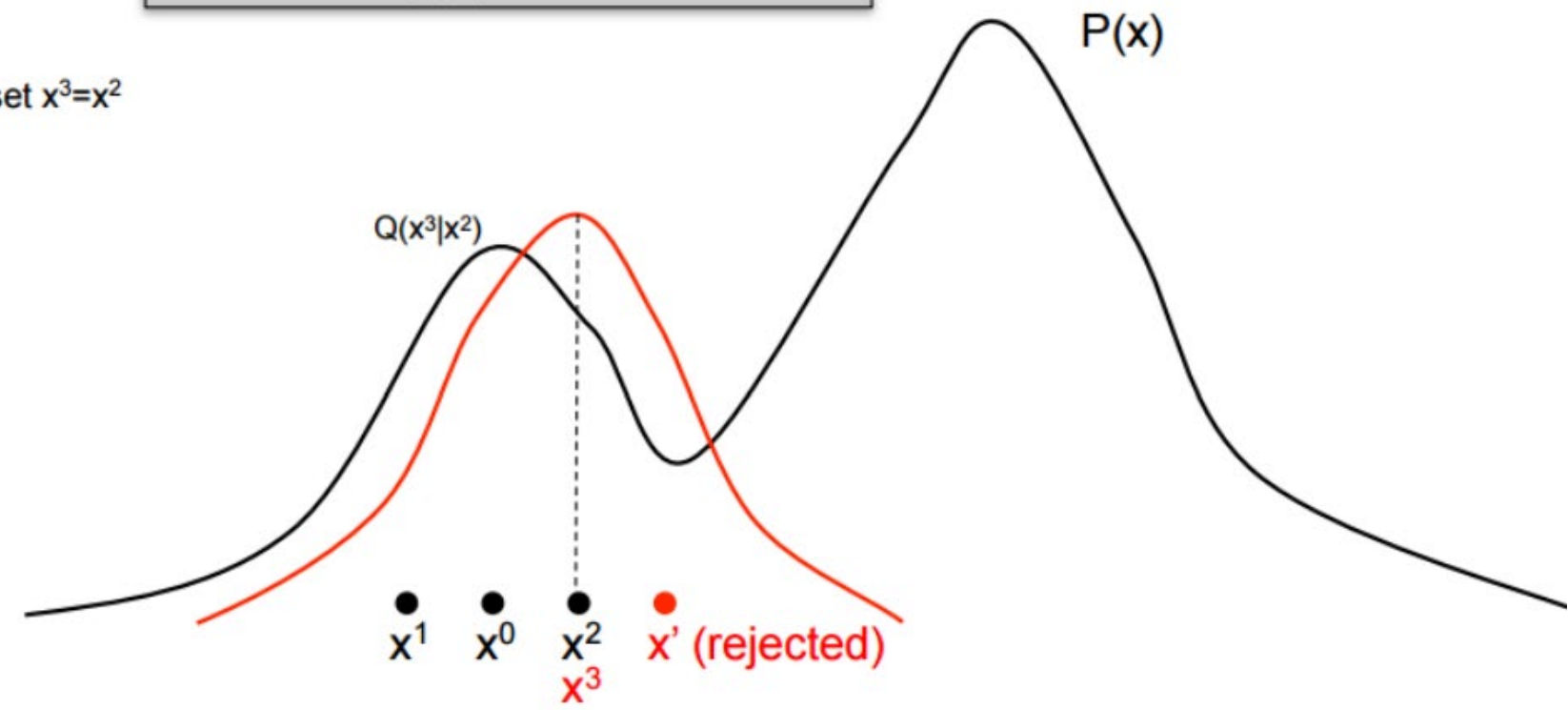


# Markov chain Monte Carlo (MCMC)

- Metropolis-Hastings example:

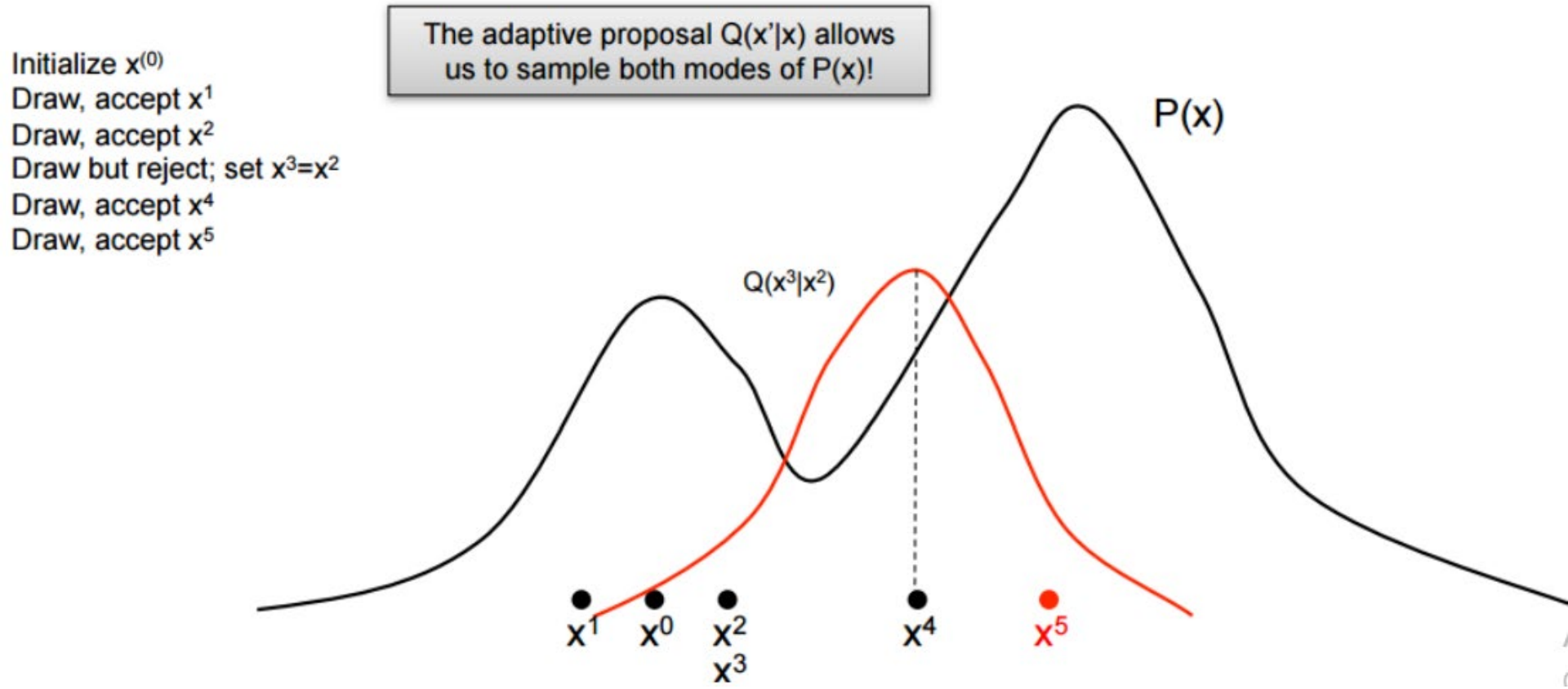
Initialize $x^{(0)}$
Draw, accept $x^1$
Draw, accept $x^2$
Draw but reject; set $x^3=x^2$

We reject because $P(x')/P(x^2)$ is very small, hence $A(x'|x^2)$ is close to zero!



# Markov chain Monte Carlo (MCMC)

- Metropolis-Hastings example:



# Next topic

- Probabilistic graphical models
  - Learning