CSE 250A. Principles of AI

Probabilistic Reasoning and Decision-Making

Lecture 7 – Inference and learning in BNs

Lawrence Saul
Department of Computer Science and Engineering
University of California, San Diego

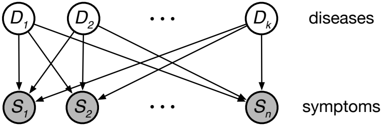
Fall 2021

Outline

Approximate inference

- **Problem (for loopy BNs)**

  Given a set E of evidence nodes, and a set Q of query nodes, how to estimate the posterior distribution $P(Q|E)$?



- **Stochastic sampling methods**

  LAST CLASS

  1. Rejection sampling — **slow**
  2. Likelihood weighting — **faster**

  TODAY

  3. Markov chain Monte Carlo (MCMC) — **fastest**

Likelihood weighting

- **Make N forward passes through the BN:**

  Sample non-evidence nodes based on values of parents.
  Fix evidence nodes to desired values.

- **For single query and evidence nodes:**

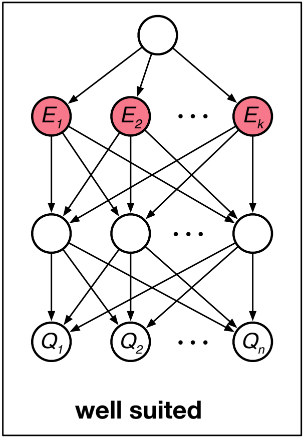
$$P(Q\!=\!q|E\!=\!e) \;\approx\; \frac{\sum_{i=1}^{N} I(q,q_i)\;\overbrace{P(E\!=\!e|\mathrm{pa}_i(E))}^{\textbf{likelihood weight}}}{\sum_{i=1}^{N} P(E\!=\!e|\mathrm{pa}_i(E))}$$

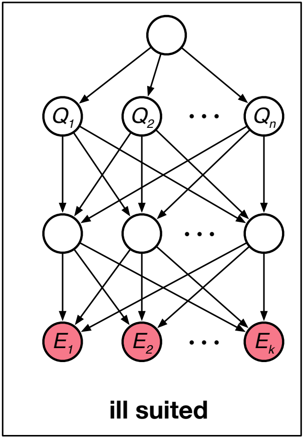
- **For multiple query and evidence nodes:**

$$P(Q\!=\!q, Q'\!=\!q'|E\!=\!e, E'\!=\!e')$$

$$\approx \quad \frac{\sum_{i=1}^{N} I(q,q_i)\, I(q',q_i')\, P(E\!=\!e|\mathrm{pa}_i(E))\, P(E'\!=\!e'|\mathrm{pa}_i(E'))}{\sum_{i=1}^{N} P(E\!=\!e|\mathrm{pa}_i(E))\, P(E'\!=\!e'|\mathrm{pa}_i(E'))}$$

Best and worst cases for likelihood weighting



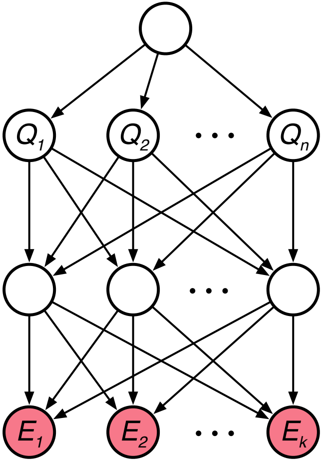


Left — rare evidence affects how query nodes are sampled.
Right — rare evidence is unlikely to occur with high probability.

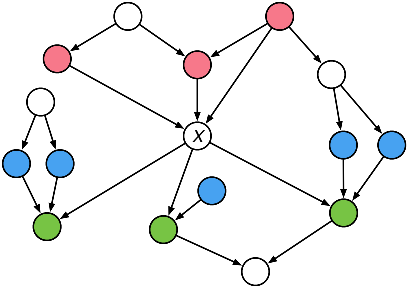
What next?

To handle this case, especially with rare evidence, we need the evidence nodes to affect how other nodes are sampled.

We need a way to sample nodes **in any order**—not only in a forward pass when they are conditioned on their parents.



Markov blanket



HW 2 reminder

- **Definition**

  The Markov blanket B_X of a node X consists of its **parents**, **children**, and **spouses** (i.e., parents of children).

- **Theorem**

  The node X is conditionally independent of **the nodes outside** its Markov blanket given **the nodes inside** its Markov blanket.

Test your understanding

Let X be a node in a belief network.
Let B_X denote its Markov blanket (i.e., parents, children, spouses).
Let Y be any node such that $Y \notin X \cup B_X$.

True or False?

1. The parents, children, and spouses of X are nonoverlapping sets of nodes.

   False. A spouse can also be a parent or a child.

2. The above sets are nonoverlapping in a polytree.

   True. A parent is never a child, and a spouse as either creates a loop.

3. $P(X|B_X, Y) = P(X|B_X)$ is only guaranteed to be true in a polytree.

   False. The theorem holds in any BN (loopy or not).

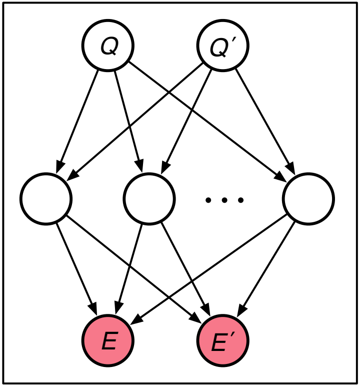
Outline

1. Review
2. Markov chain Monte Carlo
3. Learning in BNs

Approximate inference

Query nodes Q, Q'

Evidence nodes E, E'

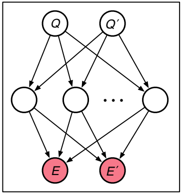


How to estimate $P(Q=q, Q'=q' | E=e, E'=e')$**?**

Markov chain Monte Carlo simulation

- **Initialization**

  Fix evidence nodes to observed values e, e'.
  Initialize non-evidence nodes to random values.



- **Repeat N times**

  Pick a non-evidence node X at random.
  Use **Bayes rule** to compute $P(X|B_X)$.
  Resample $x \sim P(X|B_X)$.
  Take a snapshot of all the nodes in the BN.

- **Estimate**

  Count the snapshots $N(q, q') \leq N$ with $Q\!=\!q$ and $Q'\!=\!q'$.

$$\boxed{P(Q\!=\!q, Q'\!=\!q'|E\!=\!e, E'\!=\!e') \;\approx\; \frac{N(q,q')}{N}}$$

Properties of MCMC

Under reasonable conditions ...

1. This sampling procedure defines an ergodic Markov chain over the non-evidence nodes of the BN.
2. The stationary distribution of this Markov chain is equal to the BN's posterior distribution over its non-evidence nodes.
3. The estimates from MCMC converge in the limit:

$$\lim_{N\to\infty} \frac{N(q,q')}{N} \;\rightarrow\; P(Q=q, Q'=q' | E=e, E'=e')$$

MCMC versus likelihood weighting (LW)

- **How they sample**

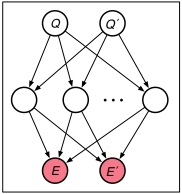
$$\left.\begin{array}{r}\textbf{LW}\\ \textbf{MCMC}\end{array}\right\}\ \text{samples non-evidence nodes from}\ \left\{\begin{array}{l}P(X|\mathrm{pa}(X))\\ P(X|B_X)\end{array}\right.$$

- **Cost per sample**

  LW can read off $P(X|\mathrm{pa}(X))$ from each CPT.
  MCMC must compute $P(X|B_X)$ before each sample.

- **Convergence**

  LW is slow for rare evidence in leaf nodes.
  MCMC can be much faster in this situation.



Outline

1. Review
2. Markov chain Monte Carlo
3. Learning in BNs

Flashback to first lecture

Outline

1. Administrivia
2. Course Overview
   - Probabilistic reasoning
   - **Learning from data**
   - Sequential modeling
   - Planning and decision-making
3. Conclusion

Learning in BNs

- **Where do BNs come from?**

  Sometimes an expert can provide the DAG and CPTs.
  But not always — especially not in very complex domains.

- **What is the alternative?**

  With sufficient data, we can estimate useful models.
  This is the central idea of *machine learning*.

- **What are some applications?**

HW 4	Language modeling
HW 5	Visual object recognition
HW 8	Recommender systems

Maximum likelihood (ML) estimation

- **Here's a simple idea:**

  Model data by the BN that assigns it the highest probability.
  In other words, choose the DAG and CPTs to **maximize**

  $$P(\text{observed data} \mid \text{DAG \& CPTs}).$$

  This probability is known as the **likelihood**.

- **But is this too simple?**

  The data may be unrepresentative or too limited.
  This is one failure mode of ML estimation.

ML Estimation in BNs

- **In CSE 250A**

  We will always assume the DAG is given.
  But we will study several cases for learning the CPTs.

Case	CPTs	data	HW
1	tabular	complete	4
2a	Gaussian	complete	4
2b	sigmoid	complete	5
3	mixed	incomplete	6-8

- **Beyond CSE 250A**

  Learning DAGs as well as CPTs
  Learning in BNs when inference is intractable
  Beyond ML estimation – Bayesian learning

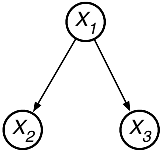
Learning with complete data and tabular CPTs

ASSUMPTIONS

1. The DAG is fixed (and known) over a finite set of discrete random variables $\{X_1, X_2, \ldots, X_n\}$.
2. The data consists of T complete (or fully observed) instantiations of all the nodes in the BN.
3. CPTs enumerate $P(X_i = x|\mathrm{pa}(X_i) = \pi)$ as lookup tables; each must be estimated for all values of x and π.

Example

- **Fixed DAG over discrete random variables**



$$\begin{aligned} X_1 &\in \{1,2,3\} \\ X_2 &\in \{1,2,3,4\} \\ X_3 &\in \{1,2,3,4,5\} \end{aligned}$$

- **Data set**

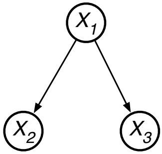
example	x_1	x_2	x_3
1	1	4	5
2	3	2	4
3	2	1	3
⋮	⋮	⋮	⋮
T	1	3	2

Note that if T is sufficiently large, some rows are destined to repeat.

We can also denote the data set as $\left\{ \left(x_1^{(t)}, x_2^{(t)}, x_3^{(t)} \right) \right\}_{t=1}^{T}$.

Example

- **Fixed DAG over discrete random variables**



$$\begin{aligned} X_1 &\in \{1,2,3\} \\ X_2 &\in \{1,2,3,4\} \\ X_3 &\in \{1,2,3,4,5\} \end{aligned}$$

- **Data set**

example	x_1	x_2	x_3
1	1	4	5
2	3	2	4
3	2	1	3
⋮	⋮	⋮	⋮
T	1	3	2

How to choose the CPTs so that the BN maximizes the probability of this data set?

ML estimation

- **IID assumption**

  The examples are assumed to be *independently and identically distributed* (IID) from the joint distribution of the BN.

- **Probability of IID data**

$$P(data) = \prod_{t=1}^{T} P\left(X_1 = x_1^{(t)}, X_2 = x_2^{(t)}, \ldots, X_n = x_n^{(t)}\right)$$

- **Probability of t^{th} example**

$$P\left(X_1 = x_1^{(t)}, X_2 = x_2^{(t)}, \ldots, X_n = x_n^{(t)}\right)$$

$$= \prod_{i=1}^{n} P\left(X_i = x_i^{(t)} \middle| X_1 = x_1^{(t)}, \ldots, X_{i-1} = x_{i-1}^{(t)}\right) \quad \boxed{\textbf{product rule}}$$

$$= \prod_{i=1}^{n} P\left(X_i = x_i^{(t)} \middle| \mathrm{pa}(X_i) = \mathrm{pa}_i^{(t)}\right) \quad \boxed{\textbf{conditional independence}}$$

Computing the log-likelihood

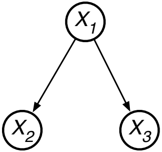
$$
\begin{aligned}
\mathcal{L} &= \log P(\mathrm{data}) \\
&= \log \prod_{t=1}^{T} P\left(x_1^{(t)}, x_2^{(t)}, \ldots, x_n^{(t)}\right) && \boxed{\textbf{IID}} \\
&= \log \prod_{t=1}^{T} \prod_{i=1}^{n} P\left(x_i^{(t)} \middle| \mathrm{pa}_i^{(t)}\right) && \boxed{\textbf{product rule \& CI}} \\
&= \sum_{t=1}^{T} \sum_{i=1}^{n} \log P\left(x_i^{(t)} \middle| \mathrm{pa}_i^{(t)}\right) && \boxed{\log pq = \log p + \log q} \\
&= \sum_{i=1}^{n} \underbrace{\sum_{t=1}^{T} \log P\left(x_i^{(t)} \middle| \mathrm{pa}_i^{(t)}\right)}_{\textbf{sum over examples}} && \boxed{\textbf{sums can be reordered}}
\end{aligned}
$$

Counting co-occurrences

- **Counts**

  Let $\text{count}(X_i = x, \text{pa}_i = \pi)$ denote the number of examples where $X_i = x$ and $\text{pa}_i = \pi$.

- **Example**



x_1	x_2	x_3
1	4	5
3	2	4
2	1	3
2	1	4
1	3	5
1	3	2

$$\begin{aligned}
\text{count}(X_1 = 1) &= 3 \\
\text{count}(X_1 = 2) &= 2 \\
\text{count}(X_1 = 3) &= 1 \\
\text{count}(X_2 = 1, X_1 = 2) &= 2 \\
\text{count}(X_2 = 3, X_1 = 1) &= 2 \\
&\vdots \\
\text{count}(X_3 = 5, X_1 = 1) &= 2
\end{aligned}$$

Note: these counts can be compiled in one pass through the data set.

Computing the log-likelihood

Next: replace the **unweighted** sum over examples at each node by a **weighted** sum over its values and those of its parents.

$$
\begin{aligned}
\mathcal{L} &= \sum_{i=1}^{n} \sum_{t=1}^{T} \log P\left(x_i^{(t)} \middle| \mathrm{pa}_i^{(t)}\right) \qquad \boxed{\textbf{unweighted}} \\
&= \sum_{i=1}^{n} \sum_{x} \sum_{\pi} \mathrm{count}(X_i = x, \mathrm{pa}_i = \pi) \log P(X_i = x | \mathrm{pa}_i = \pi)
\end{aligned}
$$

$$\boxed{\textbf{weighted}}$$

These two expressions compute the exact same sum!
But the latter has a much more appealing form ...

Interpreting the log-likelihood

$$\mathcal{L} = \sum_i \sum_x \sum_\pi \overbrace{\mathrm{count}(X_i\!=\!x, \mathrm{pa}_i\!=\!\pi)}^{\textbf{constants of the data}} \underbrace{\log P(X_i\!=\!x|\mathrm{pa}_i\!=\!\pi)}_{\textbf{CPTs to optimize}}$$

- **The log-likelihood for complete data is a triple sum over**

  i — the nodes in the BN
  x — the values of each node X_i
  π — the values π of the parents of X_i

- **How to optimize?**

  Intuitively, the larger the $\mathrm{count}(X_i\!=\!x, \mathrm{pa}_i\!=\!\pi)$,
  the larger we should choose $P(X_i\!=\!x|\mathrm{pa}_i\!=\!\pi)$.

Maximum likelihood CPTs

- **Solution without proof**

$$P_{\mathrm{ML}}(X_i = x | \mathrm{pa}_i = \pi) \;\propto\; \mathrm{count}(X_i = x, \mathrm{pa}_i = \pi)$$

- **Normalized expressions**

$$P_{\mathrm{ML}}(X_i = x | \mathrm{pa}_i = \pi) \;=\; \frac{\mathrm{count}(X_i = x, \mathrm{pa}_i = \pi)}{\mathrm{count}(\mathrm{pa}_i = \pi)} \qquad \boxed{\textbf{node with parents}}$$

$$P_{\mathrm{ML}}(X_i = x) \;=\; \frac{\mathrm{count}(X_i = x)}{T} \qquad \boxed{\textbf{root node}}$$

- **Next lecture — proof and applications ...**