10/12 **Approximate inference**

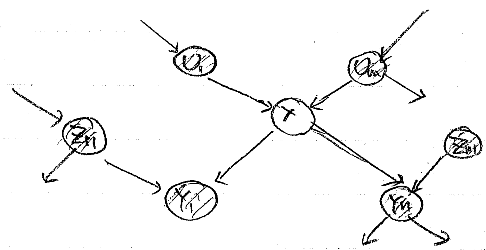
Last class : 1) Rejection sampling (slow)

2) Likelihood weighting (faster)

Today : 3) Markov chain Monte Carlo (MCMC) (fastest)

<u>Def</u> : Markov blanket B_x of node X consists of parents, children, and spouses of X.

<u>Thm</u> : X is conditionally independent of all nodes outside B_x given nodes in B_x.



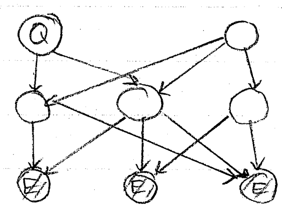
MCMC simulation

Query nodes Q.

Evidence nodes E

How to estimate $P(Q=q \mid E=e)$?

query



evidence (shaded)

* To estimate $P(Q=q \mid E=e)$:

- fix evidence nodes to observed values.
- initialize non-evidence nodes at random.
- repeat N times:
  - pick non-evidence node $X \notin E$ at random.
  - use Bayes rule to compute $P(X \mid B_x)$ where B_x is fixed to current values.
  - resample X from $P(X=x \mid B_x)$

* Count # samples $N(q)$ where $Q=q$

* Estimate $P(Q=q \mid E=e) \approx N(q)/N$.

Converges in limit $N \to \infty$ to correct answer.

* Key difference between likelihood weighting (LW) and MCMC:

LW, MCMC > sample non-evidence nodes from < $P(X \mid pa(X))$, $P(X \mid B_x)$

Learning

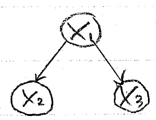
* BN = DAG + CPTs not always available from experts.
  How to learn from examples?
* Maximum likelihood (ML) estimation.
  - simplest form of learning in BNs.
  - choose ("estimate") model (DAG + CPTs) to maximize

$$\underbrace{P(\text{observed data} \mid \text{DAG} + \text{CPTs})}_{\text{likelihood.}}$$

Case I known structure of DAG, lookup tables for CPTs, complete data

- DAG is fixed over some known, finite set of discrete nodes $\{X_1, X_2, \cdots, X_n\}$
- CPTs enumerate $P(X_i = x \mid pa(X_i) = \pi)$ as lookup table
  ↖ parent configuration.
- Data is T complete instantiations of nodes in BN.

Ex:



example #	X_1	X_2	X_3
1	1	0	1
2	0	0	0
3	0	1	0
⋮			
T	1	0	1

jargon: "complete data", "fully observed", "no hidden nodes", "fully visible".

More generally, denote data as $\{(x_1^{(t)}, x_2^{(t)}, \cdots, x_n^{(t)})\}_{t=1}^{T}$.

* IID assumption
  Samples are independently identically distributed from joint distribution $P(X_1, X_2, \cdots, X_n)$ of BN.
* Probability of IID data set.

$$P(\text{data}) = \underbrace{\prod_{t=1}^{T}}_{\text{(product over rows)}} P(X_1 = x_1^{(t)}, X_2 = x_2^{(t)}, \cdots, X_n = x_n^{(t)}) \text{ due to IID assumption.}$$

Probability of t-th example:

$$P(X_1 = x_1^{(t)}, X_2 = x_2^{(t)}, \cdots, X_n = x_n^{(t)}) = \prod_{i=1}^{n} P(X_i = x_i^{(t)} \mid X_1 = x_1^{(t)}, X_2 = x_2^{(t)}, \cdots, X_{i-1} = x_{i-1}^{(t)})$$

↙ product rule

$$= \prod_{i=1}^{n} P(X_i = x_i^{(t)} \mid pa(X_i) = pa_i^{(t)})$$

↖ conditional independence from DAG.

* log-likelihood $\mathcal{L}$

$$\begin{aligned}
\mathcal{L} &= \log P(\text{DATA}) \\
&= \log \prod_{t=1}^{T} P(X_1^{(t)}, X_2^{(t)}, \cdots, X_n^{(t)}) && \text{IID} \\
&= \log \prod_{t=1}^{T} \prod_{i=1}^{n} P(X_i^{(t)} \mid pa_i^{(t)}) && \text{due to prod rule \& CI} \\
&= \sum_{t=1}^{T} \sum_{i=1}^{n} \log P(X_i^{(t)} \mid pa_i^{(t)}) \\
&= \sum_{i=1}^{n} \sum_{t=1}^{T} \log P(X_i^{(t)} \mid pa_i^{(t)}) && \text{swapping order of summation.}
\end{aligned}$$

Let count $(X_i = x, pa_i = \pi)$ denote # examples for which $X_i = x$ and $pa_i = \pi$.

$$\mathcal{L} = \sum_{i=1}^{n} \underbrace{\sum_{x}}_{\text{possible values of } X_i} \underbrace{\sum_{\pi}}_{\text{possible values of } pa(X_i)} \underbrace{\text{count}(X_i = x, pa_i = \pi)}_{\text{properties of data}} \log \underbrace{P(X_i = x \mid pa_i = \pi)}_{\text{unknowns to be optimized (learned from data)}}$$

Write $\mathcal{L} = \sum_{i,\pi} \mathcal{L}_{i\pi}$ where $\mathcal{L}_{i\pi} = \sum_{x} \text{count}(X_i = x, pa_i = \pi) \log P(X_i = x \mid pa_i = \pi)$

From decomposition $\mathcal{L} = \sum_{i,\pi} \mathcal{L}_{i\pi}$, we can independently optimize CPT entries at each node in BN and for each parent configuration of that node.

* ML estimation

For each node X_i, for each row π of CPT, maximize $\mathcal{L}_{i\pi}$ subject to:

(i) $\sum_{x} P(X_i = x \mid pa_i = \pi) = 1$

(ii) $P(X_i = x \mid pa_i = \pi) \geq 0$.

* Shorthand notation at node i, row π of CPT:

let $C_\alpha = \text{count}(X_i = \alpha, pa_i = \pi)$

let $p_\alpha = P(X_i = \alpha \mid pa_i = \pi)$

how to maximize $\sum_{\alpha} C_\alpha \log p_\alpha$ subject to $\begin{cases} \sum p_\alpha = 1 \\ p_\alpha \geq 0 \end{cases}$?

Solution : $\boxed{p_\alpha = C_\alpha \Big/ \sum_{\beta} C_\beta}$

ML Solution :
$$\begin{aligned}
P_{ML}(X_i = x \mid pa_i = \pi) &= \frac{\text{count}(X_i = x, pa_i = \pi)}{\sum_{x'} \text{count}(X_i = x', pa_i = \pi)} \\
&= \text{count}(X_i = x, pa_i = \pi) \Big/ \text{count}(pa_i = \pi).
\end{aligned}$$

* Properties

– Asymptotically correct

$P_{ML}(X_1, X_2, \cdots, X_n) \longrightarrow P(X_1, X_2, \cdots, X_n)$ as $T \to \infty$.

– Problematic in non-asymptotic regime. (of "sparse" data)

$$P_{ML}(X_i = x \mid pa_i = \pi) = \begin{cases} 0 & \text{if count}(X_i = x, pa_i = \pi) = 0 \\ \text{undefined} & \text{if count}(pa_i = \pi) = 0. \end{cases}$$

Ex: Markov models of language

* let w_ℓ denote ℓ^{th} word in sentence (or utterance or paragraph)

  How to model $P(w_1, w_2, \cdots, w_L)$?

* Simplifying assumptions

  (i) $P(w_\ell \mid w_1, w_2, \cdots, w_{\ell-1}) = P(w_\ell \mid w_{\ell-(n-1)}, w_{\ell-(n-2)}, \cdots, w_{\ell-1})$ finite context.

  (ii) $P(w_\ell = w \mid w_{\ell-(n-1)} = v_{n-1}, \cdots, w_{\ell-1} = v_1) = P(w_{\ell+1} = w \mid w_{\ell+1-(n-1)} = v_{n-1}, \cdots, w_\ell = v_1)$

  position invariance