TODAY

Approximate marginalization in junction trees (Jensen and Andersen)

Variational approximation (Jaakola and Jordan)

Brief review

All exact algorithms marginalize over local regions.

Approximate marginal on local region.

- Approximate JT, and search

Approximate regions

- Boyen & Koller, LPE [Draper]

Approximate regions and marginal

- MCMC

Approximate marginal on global net.

- Forward sampling

“Ordered” Marginalization

Sum the largest terms of marginal.

Lots of examples

- Search-based: Top-N [Henrion], Search [Poole]
  - Use best-first search to identify largest terms in the joint.
- Bounded Conditioning: [Horvitz]
  - Use prior bound on the size of cutset weights to bound contribution to the joint.
- Approximation in Junction Trees [Jensen & Andersen]

Approximating Marginalization in Cliques

Zero-compression

- Sparse array representation

D1	D2	D3	S	Potential
0	0	0	0	0.85738
0	0	0	1	0
0	0	1	0	0.02256
0	0	1	1	0.02256
0	1	0	0	0.02256
0	1	0	1	0.02256
0	1	1	0	0.00059
0	1	1	1	0.00178
1	0	0	0	0.02256
1	0	0	1	0.02256
1	0	1	0	0.00059
1	0	1	1	0.00178
1	1	0	0	0.00059
1	1	0	1	0.00178
1	1	1	0	1.6E-05
1	1	1	1	0.00011

→

D1	D2	D3	S	Potential
0	0	0	0	0.85738
0	0	1	0	0.02256
0	0	1	1	0.02256
0	1	0	0	0.02256
0	1	0	1	0.02256
0	1	1	0	0.00059
0	1	1	1	0.00178
1	0	0	0	0.02256
1	0	0	1	0.02256
1	0	1	0	0.00059
1	0	1	1	0.00178
1	1	0	0	0.00059
1	1	0	1	0.00178
1	1	1	0	1.6E-05
1	1	1	1	0.00011

Approximation

Set threshold (0.001)

Set all potentials below this threshold to 0

D1	D2	D3	S	Potential
0	0	0	0	0.85738
0	0	1	0	0.02256
0	0	1	1	0.02256
0	1	0	0	0.02256
0	1	0	1	0.02256
0	1	1	0	0.00059
0	1	1	1	0.00178
1	0	0	0	0.02256
1	0	0	1	0.02256
1	0	1	0	0.00059
1	0	1	1	0.00178
1	1	0	0	0.00059
1	1	0	1	0.00178
1	1	1	0	1.6E-05
1	1	1	1	0.00011

→

D1	D2	D3	S	Potential
0	0	0	0	0.85738
0	0	1	0	0.02256
0	0	1	1	0.02256
0	1	0	0	0.02256
0	1	0	1	0.02256
0	1	1	1	0.00178
1	0	0	0	0.02256
1	0	0	1	0.02256
1	0	1	1	0.00178
1	1	0	1	0.00178
	SUM			0.99809375

$$bounds = \left[\frac{P\{X,E\}}{P\{E\}-\delta}, \frac{P\{X,E\}-\delta}{P\{E\}-\delta} \right]$$

Variational Approximation

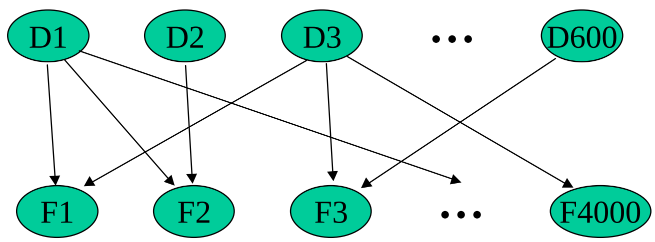
Sketch:

1. Calculate formula for noisy-or
2. Show that negative evidence noisy-or nodes always factor
3. Develop approximation for positive evidence

   Upper bound using convex analysis

   Lower bound using Jensen’s inequality

Inspiration: QMR-DT

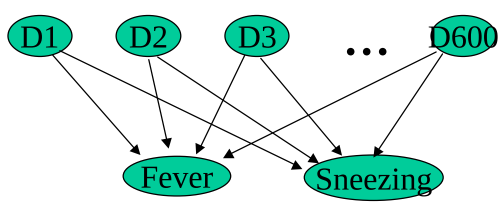


On standard CPC cases: Median clique size 151.5
BN20: All nodes are binary

Intuition

“Many of the sums (marginals) in QMR-DT are sums over large numbers of random variables.

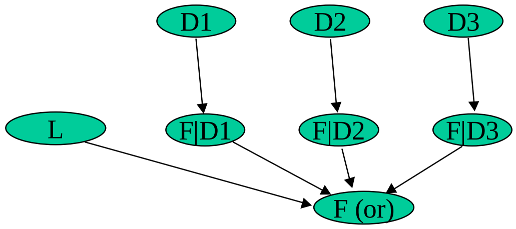
Laws of large numbers suggest that these sums may yield predictable numeric results over the ensemble of their summands”



Especially when high leaks, lots of parents.

Correlations not important

Noisy-Or



Distinguished state is

$$d = 0, \; f = 0$$

Remember:

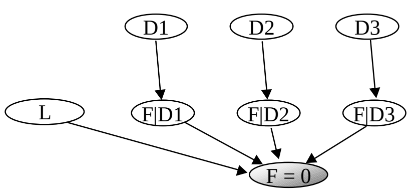
$$P\{f_i = 0 \mid d_j = 0\} = 1$$

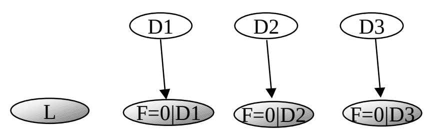
What is P(f=0|D)?

$$P\{f_i = 0 \mid D\} = P\{f_i = 0 \mid L\} \prod_{d_j \in \mathrm{pa}(f_i)} P\{f_i = 0 \mid d_j\}$$

Factoring Negative Evidence

$$P\{f, D\} = P\{D\} P\{f^- \mid D\} P\{f^+ \mid D\}$$





$$P\{D\}P\{f^- \mid D\} = \prod_{d_i} \left(P\{d_i\} \prod_{f_j \in ch(d_i) \cap f^-} P\{f_j = 0 \mid d_i\} \right) \left(\prod_{f_j \in f^-} P\{f_j = 0 \mid L\} \right)$$

Function of d_i only — constant

So... d's are independent after observing negative evidence

“Absorb” Negative Findings

$$P\{f,D\} \propto \left(\prod_{d_i \in D} P\{d_i \mid f^-\} \right) P\{f^+ \mid D\}$$

$$P\{d_i\} \longleftarrow P\{d_i \mid f^-\}$$

Hard part is factoring positive findings

$$P\{f_i = 1 \mid D\} = 1 - P\{f_i = 0 \mid L\} \prod_{d_j \in \mathrm{pa}(f_i)} P\{f_i = 0 \mid d_j\}$$

OK, Where are we going?

1. Rewrite noisy-or expression for positive evidence
2. Find an approximation (called a variational approximation) that factors.
   - Upper bound: convex analysis
   - Lower bound: Jensen’s inequality

1. Rewrite Noisy Or

$$P\{f_i = 0 \mid D\} = P\{f_i = 0 \mid L\} \prod_{d_j \in \mathrm{pa}(f_i)} P\{f_i = 0 \mid d_j\}$$

Let $P\{f_i = 1 \mid L\} = q_{i0}$

$$P\{f_i = 1 \mid d_j = 1\} = q_{ij}$$

$$P\{f_i = 0 \mid D\} = (1 - q_{i0}) \prod_{d_j \in pa(f_i)} (1 - q_{ij})^{d_j}$$

$$P\{f_i = 1 \mid D\} = 1 - \left[(1 - q_{i0}) \prod_{d_j \in pa(f_i)} (1 - q_{ij})^{d_j} \right]$$

1. Transform a bit more...

$$P\{f = 1_i \mid D\} \equiv P\{f_i^+ \mid D\}$$

$$= 1 - (1 - q_{i0}) \prod_{d_j \in pa(f_i)} (1 - q_{ij})^{d_j}$$

Let $\theta_{ij} = -\ln(1 - q_{ij})$

$$P\{f_i^+ \mid D\} = 1 - e^{-\theta_{i0} - \sum_{d_j \in pa(f_i)} \theta_{ij} d_j}$$

Variational Transform

(Convex analysis)

Any concave function f(x) can be represented as

$$f(x) = \min_{\xi}\left[\xi^T x - f^*(\xi)\right] \quad \text{(variational transform)}$$

$$f^*(\xi) = \min_{x}\left[\xi^T x - f(x)\right]$$

The variational approximation (an *upper* bound):

$$f(x) \le \xi^T x - f^*(\xi)$$

This approximation is *linear* in *x*!

Variational Upper Bound

$$P\left\{f_i^+ \mid D\right\} = 1 - e^{-\theta_{i0} - \sum_{d_j \in pa(f_i)} \theta_{ij} d_j} = e^{\log\left(1 - e^{-x}\right)}$$

Let $f(x) = e^{\log\left(1 - e^{-x}\right)}$

$$f^*(\xi) = \min_{x}\left[\xi x - \log\left(1 - e^{-x}\right)\right]$$

$$\frac{\partial}{\partial x}\left(\xi x - \log\left(1 - e^{-x}\right)\right) = \xi - \frac{e^{-x}}{1 - e^{-x}} = 0$$

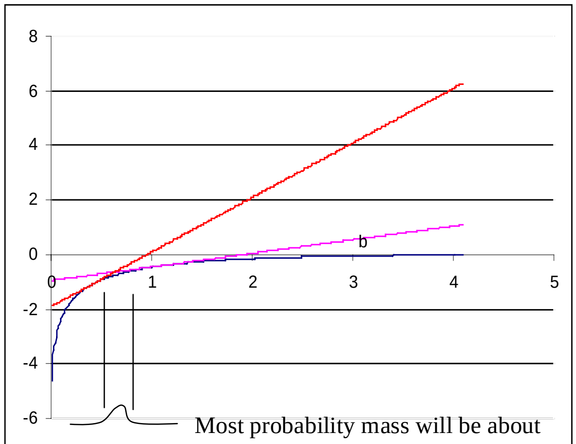
$$x = -\ln \xi + \ln(1 + \xi)$$

$$f^*(\xi) = \xi\left(-\ln \xi + \ln(1+\xi)\right) - \log\left(1 - e^{\ln \xi - \ln(1+\xi)}\right)$$

$$f^*(\xi) = -\xi \ln \xi + (1+\xi)\ln(1+\xi)$$

$$f(x) \le \xi x + \xi \ln \xi - (1+\xi)\ln(1+\xi)$$

Upper Bounds



Variational Upper Bound

$$P\{f_i^+ \mid D\} = e^{f\left(\theta_{i0} + \sum_{d_j \in pa(f_i)} \theta_{ij} d_j\right)}$$

$$\leq e^{\xi_i\left(\theta_{i0} + \sum_{d_j \in pa(f_i)} \theta_{ij} d_j\right) - f^*(\xi_i)}$$

$$\equiv P\{f_i^+ \mid D, \xi_i\}$$

This is why this approximation is cool:

$$P\{f_i^+ \mid D, \xi_i\} = e^{\xi_i \theta_{i0} - f^*(\xi_i)} \prod_j \left(e^{\xi_i \theta_{ij}}\right)^{d_j}$$

Variational Lower Bound

Jensen's Inequality

$$f\left(a+\sum_j b_j z_j\right) \geq \sum_j b_j f\left(a+z_j\right)$$

where $f(\ )$ is concave

$$\sum_j b_j = 1,$$

$$b_j > 0$$

Variational Lower Bounds

$$P\left\{f_i^+ \mid d\right\} = e^{f\left(\theta_{i0}+\sum_j \theta_{ij} d_j\right)}$$

$$\geq e^{\sum_j q_{j|i} f\left(\theta_{i0}+\frac{\theta_{ij} d_j}{q_{j|i}}\right)}$$

$$\geq e^{\sum_j q_{j|i}\left[d_j f\left(\theta_{i0}+\frac{\theta_{ij}}{q_{j|i}}\right)+\left(1-d_j\right) f\left(\theta_{i0}\right)\right]}$$

$$= e^{\sum_j q_{j|i} d_j\left[f\left(\theta_{i0}+\frac{\theta_{ij}}{q_{j|i}}\right)-f\left(\theta_{i0}\right)\right]+f\left(\theta_{i0}\right)} \equiv P\left\{f_i^+ \mid d, q_{\cdot|i}\right\}$$

The Algorithm

Some set of positive findings is computed exactly.

The rest are approximated using variational upper and lower bounds.

Optimize bounds

$$\max_q P\{f^+, d, q\} \qquad \text{via EM}$$

$$\min_\xi P\{f^+, d, \xi\} \qquad \text{via convex optimization}$$

Use optimized variational approximation to compute interesting marginals.

Intervals on marginals

$$P\{f^+, d_j\} = \sum_{D \backslash d_j} P\{f^+ \mid D\} P\{D\} \leq \sum_{D \backslash d_j} P\{f^+ \mid D, \xi\} P\{D\} \equiv P\{f^+, d_j \mid \xi\}$$

$$P\{f^+, d_j\} = \sum_{D \backslash d_j} P\{f^+ \mid D\} P\{D\} \geq \sum_{D \backslash d_j} P\{f^+ \mid D, q\} P\{D\} \equiv P\{f^+, d_j \mid q\}$$

[Draper+Hanks, 94]:

$$\frac{P\{f^+, d_j \mid q\}}{P\{f^+, \neg d_j \mid \xi\} + P\{f^+, d_j \mid q\}} \leq P\{d_j \mid f^+\} \leq \frac{P\{f^+, d_j \mid \xi\}}{P\{f^+, \neg d_j \mid q\} + P\{f^+, d_j \mid \xi\}}$$

Where we are at:

What we can do:

- Compute and approximate discrete distributions.
- (Exact computation with select continuous distribution types.)
- Compute moments on continuous distributions when problem is “friendly”
  - High relative likelihood of evidence (likelihood weighting)
  - High rate of mixing (Markov chain)

What we cannot do:

- Closed form inference on most continuous distributions.
- Approximate inference when there is a lot of evidence.