Graphical probabilistic models – inference

Jiří Kléma

Department of Computers,
FEE, CTU at Prague

http://cw.felk.cvut.cz/wiki/courses/ae4m33rzn/start

Agenda

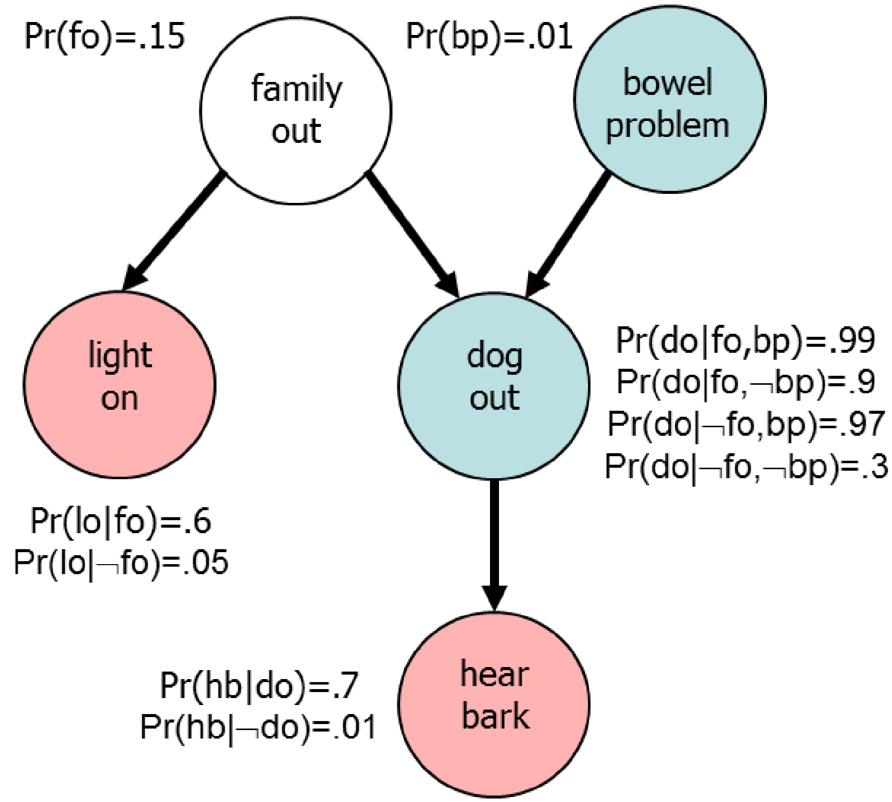
- Bayesian networks
  - fundamental tasks,
- exact **inference** and its complexity
  - straightforward enumeration
    - easy to understand but inefficient – computes joint probabilities,
    - descends to the level of atomic events,
  - acceleration by variable elimination,
- exact $\times$ approximate algorithms,
  - rejection sampling,
  - likelihood weighting,
  - Gibbs sampling,

Bayesian networks – fundamental tasks

- inference – reasoning, deduction
  - from observed events assumes on probability of other events,
  - observations (**E** – a set of evidence variables, **e** – a particular event),
  - target variables (**Q** – a set of query variables, Q – a particular query variable),
  - $Pr(\mathbf{Q}|\mathbf{e})$, resp. $Pr(Q \in \mathbf{Q}|\mathbf{e})$ to be found,
  - network is known (both graph and CPTs),
- learning network parameters from data
  - network structure (graph) is given,
  - "only" quantitative parameters (CPTs) to be optimized,
- learning network structure from data
  - propose an optimal network structure
    - which edges of the fully connected graph shall be employed?,
  - too many arcs → complicated model,
  - too few arcs → inaccurate model.

Probabilistic network – inference by enumeration

- Let us observe the following events:
  - no barking heard,
  - the door light is on.
- What is the prob of family being out?
  - searching for $Pr(fo|lo, \neg hb)$.
- Will observation influence the target event?
  - light on supports departure hypothesis,
  - no barking suggests dog inside,
  - the dog is in house when it is
    - rather healthy,
    - the family is at home.



Probabilistic network – inference by enumeration

- inference by enumeration
  - conditional probs calculated by summing the elements of joint probability table,
- how to find the joint probabilities (the table is not given)?
  - BN definition suggests:

    $Pr(FO, BP, DO, LO, HB) =$

    $= Pr(FO)Pr(BP)Pr(DO|FO, BP)Pr(LO|FO)Pr(HB|DO)$
- answer to the question?
  - conditional probability definition suggests:

    $Pr(fo|lo, \neg hb) = \frac{Pr(fo,lo,\neg hb)}{Pr(lo,\neg hb)}$
  - by joint prob marginalization we get:

    $Pr(fo, lo, \neg hb) = \sum_{BP,DO} Pr(fo, BP, DO, lo, \neg hb)$

    $Pr(fo, lo, \neg hb) = Pr(fo, bp, do, lo, \neg hb) + Pr(fo, bp, \neg do, lo, \neg hb) +$
    $+ Pr(fo, \neg bp, do, lo, \neg hb) + Pr(fo, \neg bp, \neg do, lo, \neg hb) = .15 \times .01 \times .99 \times .6 \times .3 + .15 \times$
    $.01 \times .01 \times .6 \times .99 + .15 \times .99 \times .9 \times .6 \times .3 + .15 \times .99 \times .1 \times .6 \times .99 = .033$

    $Pr(lo, \neg hb) = Pr(fo, lo, \neg hb) + Pr(\neg fo, lo, \neg hb) = .066$

Probabilistic network – inference by enumeration

- after substitution:
  $Pr(fo|lo, \neg hb) = \frac{Pr(fo,lo,\neg hb)}{Pr(lo,\neg hb)} = \frac{.033}{.066} = 0.5$
- posterior probability $Pr(fo|lo, \neg hb)$ higher than prior $Pr(fo) = 0.15$.

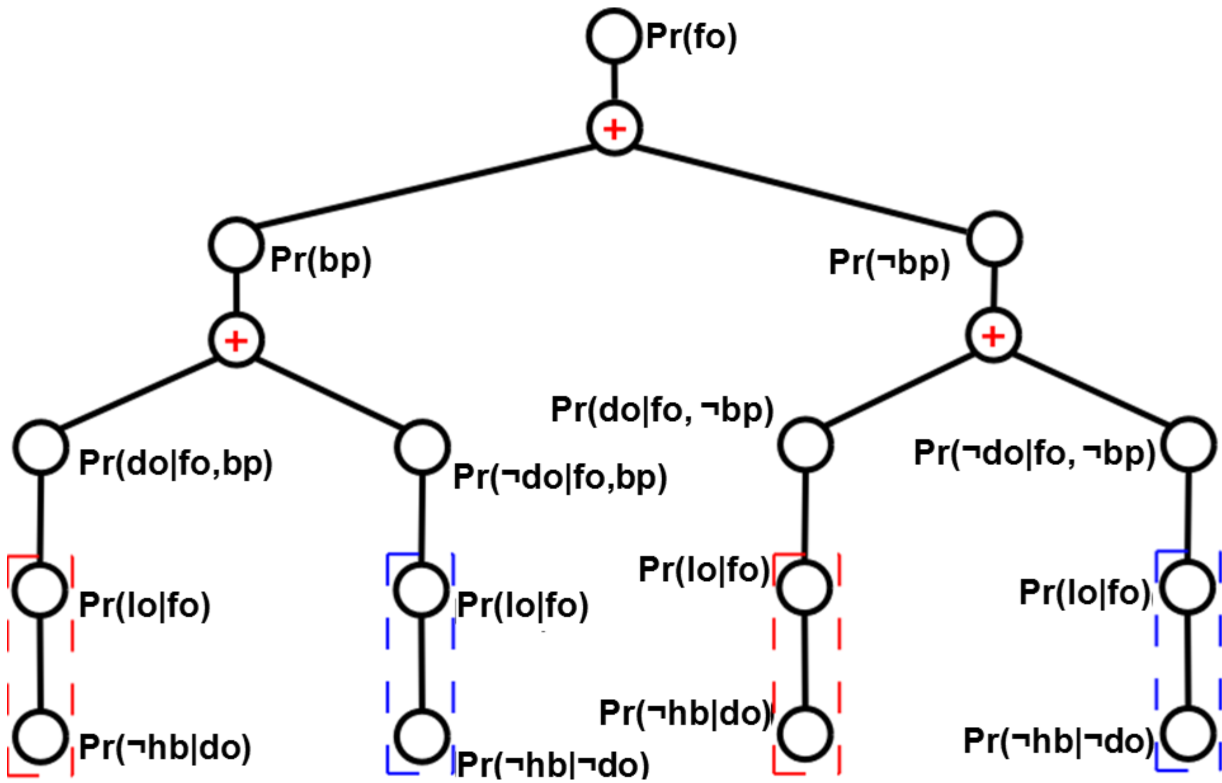
■ can we assume on complexity?

- instead of $2^5 - 1$=31 probs (either conditional or joint) 10 needed only,
- however, joint probs enumerated to answer the query
  - ∗ inference remains a NP-hard problem,
- moving summations left-to-right makes a difference, but not a principal one
  - ∗ see the evaluation tree on the next slide,

$$Pr(fo, lo, \neg hb) = \sum_{BP,DO} Pr(fo, BP, DO, lo, \neg hb) =$$
$$= Pr(fo) \sum_{BP} Pr(BP) \sum_{DO} Pr(DO|fo, BP) Pr(lo|fo) Pr(\neg hb|DO)$$

- inference by enumeration is an intelligible, but inefficient procedure,
- solution: minimize recomputations, special network types or approximate inference.

Inference by enumeration – evaluation tree



- Complexity: time $\mathcal{O}(n2^d)$, memory $\mathcal{O}(n)$
  - n . . . the number of variables, e . . . the number of evidence variables, d=n-e,
- resource of inefficiency: recomputations ($Pr(lo|fo) \times Pr(\neg hb|DO)$ for each BP value)
  - variable ordering makes a difference – $Pr(lo|fo)$ shall be moved forward.

Inference by enumeration – straightforward improvements

- variable elimination procedure

1. pre-computes factors to remove the inefficiency shown in the previous slide
   - factors serve for recycling the earlier computed intermediate results,
   - some variables are eliminated by summing them out,

     $\sum_P f_1 \times \cdots \times f_k = f_1 \times \cdots \times f_i \times \sum_P f_{i+1} \times \cdots \times f_k = f_1 \times \cdots \times f_i \times f_{\bar{P}}$,
     assumes that $f_1, \ldots, f_i$ do not depend on P,

     when multiplying factors, the pointwise product is computed
     $f_1(x_1, ..., x_j, y_1, ..., y_k) \times f_2(y_1, ..., y_k, z_1, ..., z_l) = f(x_1, ..., x_j, y_1, ..., y_k, z_1, ..., z_l)$

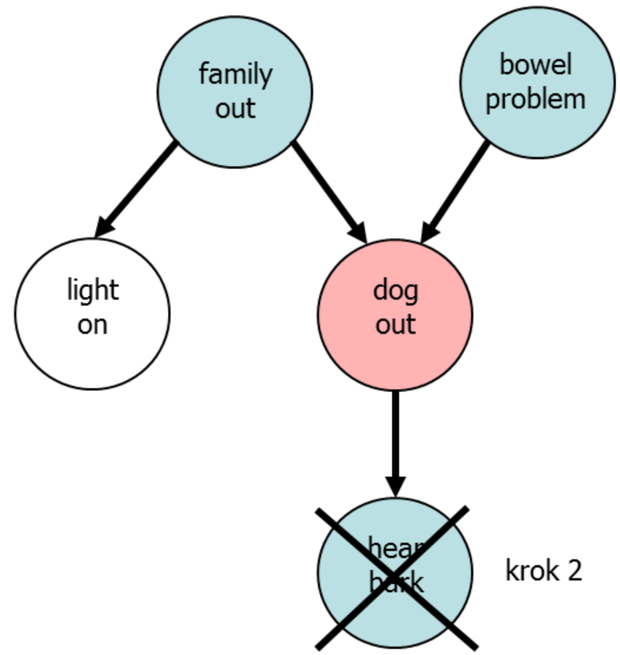
     eventual enumeration over P_1 variable, which takes all (two) possible values
     $f_{\bar{P}_1}(P_2, ..., P_k) = \sum_{P_1} f_1(P_1, P_2, ..., P_k)$,

   - execution efficiency is influenced by the variable ordering when computing,
     (finding the best order is NP-hard problem, can be optimized heuristically too),

Inference by enumeration – straightforward improvements

- **variable elimination** procedure

  2. does not consider variables irrelevant to the query
     - all the leaves that are neither query nor evidence variable,
     - the rule can be applied recursively.



- example: $Pr(lo|do)$
  - what is prob that the door light is shining if the dog is in the garden?
  - we will enumerate $Pr(LO, do)$, since:

$$Pr(lo|do) = \frac{Pr(lo,do)}{Pr(do)} = \frac{Pr(lo,do)}{Pr(lo,do)+Pr(\neg lo,do)}$$

Inference by enumeration – variable elimination

- HB is irrelevant to the particular query, why?

$\sum_{HB} Pr(HB|do) = 1$

$$Pr(LO, do) = \sum_{FO,BP,HB} Pr(FO)Pr(BP)Pr(do|FO, BP)Pr(LO|FO)Pr(HB|do) =$$
$$= \sum_{FO} Pr(FO)Pr(LO|FO) \sum_{BP} Pr(BP)Pr(do|FO, BP) \sum_{HB} Pr(HB|do)$$

- after omitting the last invariant, **factorization** may take place

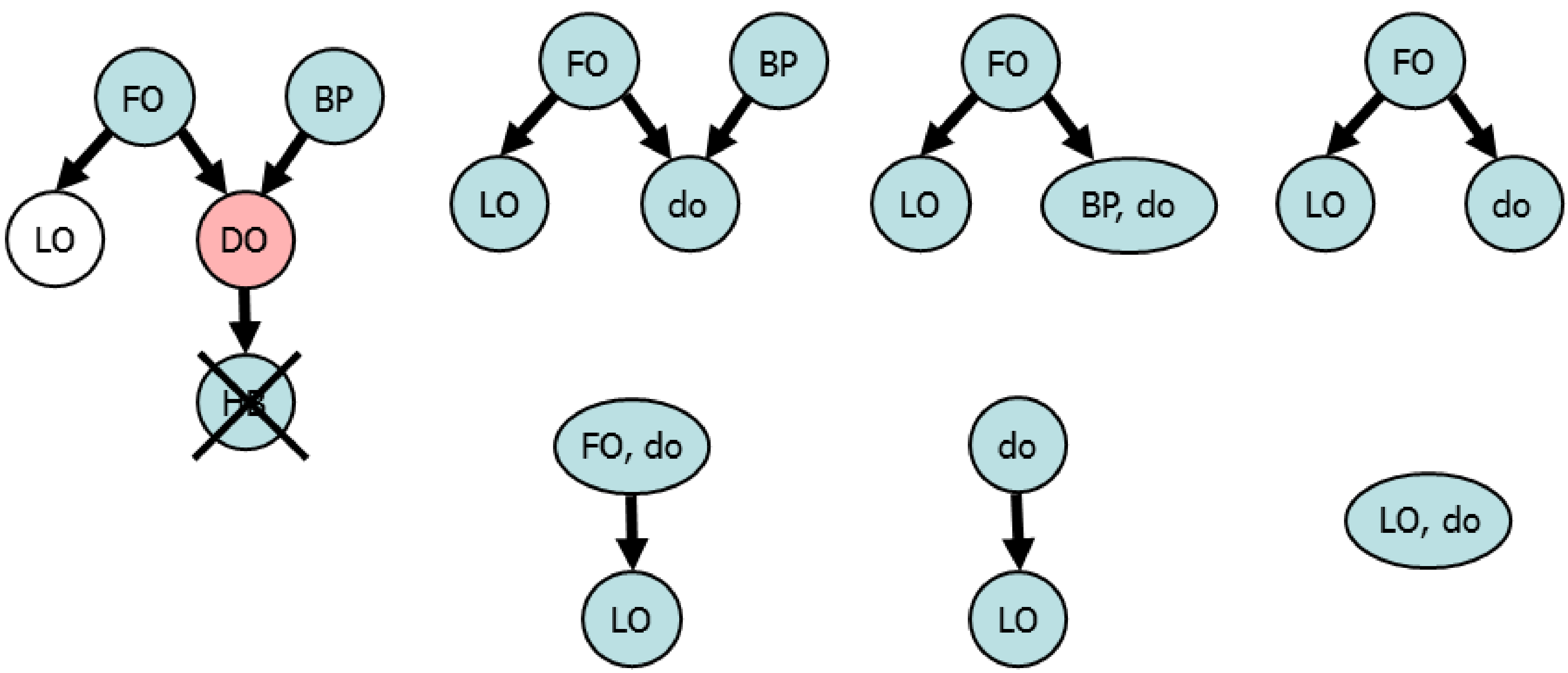
$$Pr(LO, do) = \sum_{FO} Pr(FO)Pr(LO|FO) \sum_{BP} Pr(BP)Pr(do|FO, BP) =$$
$$= \sum_{FO} Pr(FO)Pr(LO|FO) f_{\overline{BP}}(do|FO) = \sum_{FO} f_{\overline{BP},do}(FO)Pr(LO|FO) =$$
$$= f_{\overline{FO},\overline{BP},do}(LO)$$

- having the last factor (a table of two elements), one can read

$$Pr(lo|do) = \frac{f_{\overline{FO},\overline{BP},do}(lo)}{f_{\overline{FO},\overline{BP},do}(lo)+f_{\overline{FO},\overline{BP},do}(\neg lo)} = \frac{0.0941}{0.0941+0.3017} = \frac{0.0941}{0.3958} = 0.24$$

Variable elimination – factor computations

- factors are enumerated from CPTs by summing out variables
  - sum out BP: $CPT(DO)$ & $CPT(BP) \rightarrow f_{\overline{BP}}(do|FO)$
  - reformulate into: $CPT(FO)$ & $f_{\overline{BP}}(do|FO) \rightarrow f_{\overline{BP},do}(FO)$
  - sum out FO: $f_{\overline{BP},do}(FO)$ & $CPT(LO) \rightarrow f_{\overline{FO},\overline{BP},do}(LO)$



Variable elimination – factor computations

BP	$Pr(BP)$
T	0.01
F	0.99

$\times$

FO	BP	$Pr(do\|FO,BP)$
T	T	0.99
T	F	0.9
F	T	0.97
F	F	0.3

$\Rightarrow$

FO	$f_{\overline{BP}}(do\|FO)$
T	0.9009=0.99×0.01+0.9×0.99
F	0.3067=0.97×0.01+0.99×0.3

FO	$Pr(FO)$
T	0.15
F	0.85

$\times$

FO	$f_{\overline{BP}}(do\|FO)$
T	0.9009
F	0.3067

$\Rightarrow$

FO	$f_{\overline{BP},do}(FO)$
T	0.1351=0.15×0.9009
F	0.2607=0.85×0.3067

FO	$f_{\overline{BP},do}(FO)$
T	0.1351
F	0.2607

$\times$

LO	FO	$Pr(LO\|FO)$
T	T	0.6
T	F	0.05
F	T	0.4
F	F	0.95

$\Rightarrow$

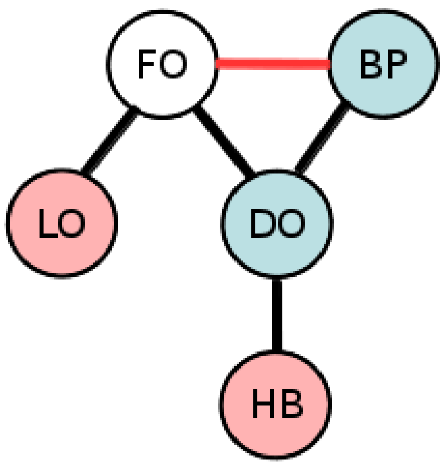
LO	$f_{\overline{FO},\overline{BP},do}(LO)$
T	0.0941=0.1351×0.6+0.2607×0.05
F	0.3017=0.1351×0.4+0.2607×0.95

Inference by enumeration – comparison of the number of operations

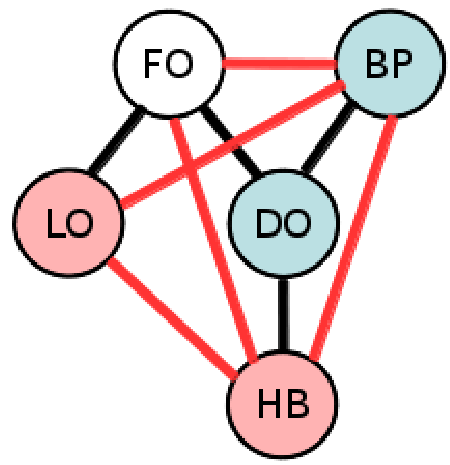
- let us take the last example
  - namely the total number of sums and products in $Pr(LO, do)$,
  - (the final $Pr(lo|do)$ enumeration is identical for all procedures),
- naïve enumeration, no evaluation tree
  - 4 products (5 vars) $\times 2^4$ ($\#$ atomic events on unevidenced variables) $+ 2^4 - 2$ sums,
  - in total 78 operations,
- using evaluation tree and a proper reordering of variables
  - takes the ordering $Pr(LO, do) = \sum_{FO} Pr(FO)Pr(LO|FO) \sum_{BP} Pr(BP)Pr(do|FO, BP) \sum_{HB} Pr(HB|do)$
  - in total 38 operations,
- with variable elimination on top of that
  - in total 14 operations (6 in Tab1, 2 in Tab2, 6 in Tab3).

Variable elimination – efficiency in general

- Given by the network structure and the variable ordering
  - exponential in the size of the largest clique in the **induced** graph,
  - somewhere between linear and NP-hard,
- induced graph
  - undirected graph, the edge exists if two variables both appear in some intermediate factor induced by the given variable ordering,



$HB \prec BP \prec LO \prec FO \prec DO$



$DO \prec FO \prec LO \prec HB \prec BP$

Variable elimination – variable ordering

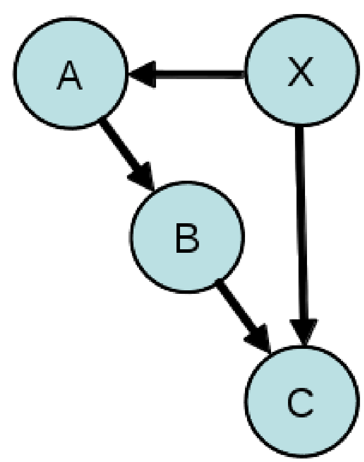
- minimize the number of **fill edges** in induced graph
  - edges introduced in the elimination step,
- NP-hard problem in general
  - greedy local methods often find near-optimal solution,
  - min-fill heuristic
    - vertex cost is the number of edges added to the graph due to its elimination,
  - always take the node that minimizes the heuristic, no backtrack.
- Step 1:

$$Pr(FO,\ldots,HB) = f_{FO}(FO)f_{BP}(BP)f_{DO}(DO,FO,BP)f_{LO}(LO,FO)f_{HB}(HB,DO)$$

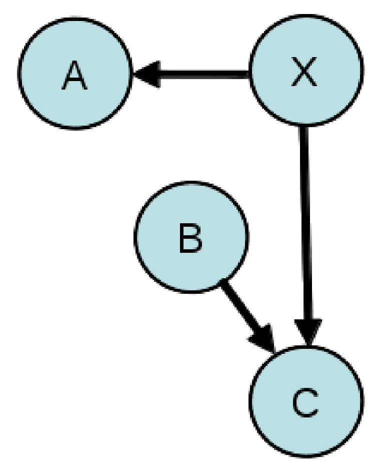
var	intermediate factor	min-fill
FO	$f_{FO}(FO)f_{DO}(DO,FO,BP)f_{LO}(LO,FO)$	3
BP	$f_{BP}(BP)f_{DO}(DO,FO,BP)$	1
DO	$f_{DO}(DO,FO,BP)f_{HB}(HB,DO)$	3
LO	$f_{LO}(LO,FO)$	0
HB	$f_{HB}(HB,DO)$	0

Semantics of factors

- Factors
  - multidimensional arrays (the same as CPTs),
  - often correspond to marginal or conditional probabilities,
  - initialized with CPTs,
  - some intermediate factors differ from any probability in the network
    - eliminate X from the left network,
    - the resulting factor does not agree with any prob in the left network,
    - it gives a conditional prob in the right network.



$f(A, B, C) = \sum_X Pr(X)Pr(A|X)Pr(C|B, X)$



$Pr(A, C|B)$

Approximate inference by stochastic sampling

- a general **Monte-Carlo** method, samples from the joint prob distribution,
- estimates the target conditional probability (query) from a sample set,
- the joint prob distribution is not explicitly given, its factorization is available only (network),

- the most straightforward is direct **forward sampling**
  1. topologically sort the network nodes
     - for every edge it holds that parent comes before its children in the ordering,
  2. instantiate variables along the topological ordering
     - take $Pr(P_j|parents(P_j))$, randomly sample P_j,
  3. repeat step 2 for all the samples (the sample size M is given a priori),
- from samples to probabilities?
  - $Pr(q|\mathbf{e}) \approx \frac{N(q,\mathbf{e})}{N(\mathbf{e})}$
  - samples that contradict evidence not used,
  - forward sampling gets inefficient if $Pr(\mathbf{e})$ is small,

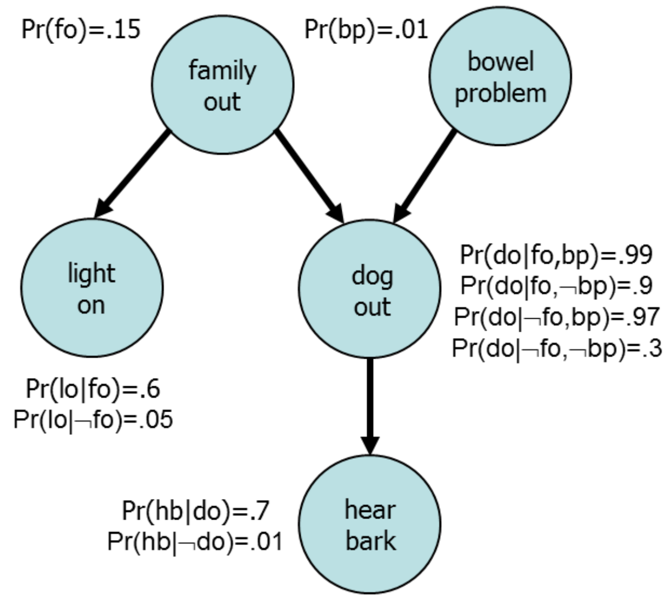
Improved stochastic sampling

- **rejection sampling**
  - rejects partially generated samples as soon as they violate the evidence event $\mathbf{e}$,
  - sample generation may stop early $\rightarrow$ slight improvement,
- **likelihood weighting**
  - avoids necessity to reject samples,
  - the values of $\mathbf{E}$ fixed, the rest of variables sampled only,
  - however, not all events are equally probable, samples must be weighted,
  - the weight equals to the likelihood of the event given the evidence,
- **Gibbs sampling**
  - the previous methods are importance sampling,
  - GS is a **Markov chain** method – the next state depends purely on the current state,
    - state = sample, generates dependent samples!
    - as it is a **Monte-Carlo** method as well $\rightarrow$ MCMC,
  - efficient sampling method namely when some of BN variable states are known
    - it again samples nonevidence variables only, the samples never rejected.

Rejection sampling – example

- FAMILY example, estimate $Pr(fo|lo, \neg hb)$
  1. topologically sort the network nodes
     - e.g., $\langle FO, BP, LO, DO, HB\rangle$ (or $\langle BP, FO, DO, HB, LO\rangle$, etc.)
  2. instantiate variables along the topological ordering
     - $Pr(FO) \rightarrow \neg fo$, $Pr(BP) \rightarrow \neg bp$,
       $Pr(LO|\neg fo) \rightarrow lo$, $Pr(DO|\neg fo, \neg bp) \rightarrow \neg do$, $Pr(HB|\neg do) \rightarrow \neg hb$
     - sample agrees with the evidence $\mathbf{e} = lo \wedge \neg hb$, no rejection needed,
  3. generate 1000 samples, repeat step 2,

- let $N(fo, lo, \neg hb)$ is 491 (the number of samples with the given values of three variables under consideration),
- in rejection sampling $N(\mathbf{e})$ necessarily equals M,
  - $Pr(fo|lo, \neg hb) \approx \frac{N(q,\mathbf{e})}{N(\mathbf{e})} = \frac{491}{1000} = 0.491$



Likelihood weighting – details

- sampling process:

  $\forall$ samples $p^m = \{P_1 = p_1^m, \ldots, P_n = p_n^m\}$, $m \in \{1, \ldots, M\}$

  1. $w^m \leftarrow 1$ (initialize the sample weight)
  2. $\forall j \in \{1, \ldots, n\}$ (instantiate variables along the topological ordering)
     - if $P_j \in \mathbf{E}$ then take p_j^m from $\mathbf{e}$ and $w^m \leftarrow w^m \times Pr(P_j|parents(P_j))$,
     - otherwise randomly sample p_j^m from $Pr(P_j|parents(P_j))$,

- from samples to probabilities?
  - evidence holds in all samples (by definition),
  - weighted averaging is applied to find $Pr(Q = P_i|\mathbf{e})$

$$Pr(p_i|\mathbf{e}) \approx \frac{\sum_{m=1}^{M} w^m \delta(p_i^m, p_i)}{\sum_{m=1}^{M} w^m} \quad \delta(i,j) = \begin{cases} 1 & \text{for } i = j \\ 0 & \text{for } i \neq j \end{cases}$$

- nevertheless, samples may have very low weights
  - turns out inefficient in large networks with evidences occuring late in the ordering.

Likelihood weighting – example

- let us approximate $Pr(fo|lo, \neg hb)$ (its exact value computed earlier is 0.5),

	p^1	p^2	p^3	...
FO	**F**	**F**	T	
BP	**F**	**F**	F	
LO	**T**	**T**	**T**	
DO	**F**	**T**	T	
HB	**F**	**F**	**F**	
w	.0495	.015	.18	

FO^1: $Pr(fo) = .15 \rightarrow \neg fo$ sampled
BP^1: $Pr(bp) = .01 \rightarrow \neg bp$ sampled
LO^1: evidence $\rightarrow lo \wedge w^1 = Pr(lo|\neg fo) = .05$
DO^1: $Pr(do|\neg fo, \neg bp) = .3 \rightarrow \neg do$ sampled
HB^1: evidence $\rightarrow \neg hb \wedge w^1 = .05 \times Pr(\neg hb|\neg do) = .0495$

FO^2: $Pr(fo) = .15 \rightarrow \neg fo$ sampled
BP^2: $Pr(bp) = .01 \rightarrow \neg bp$ sampled
LO^2: evidence $\rightarrow lo \wedge w^1 = Pr(lo|\neg fo) = .05$
DO^2: $Pr(do|\neg fo, \neg bp) = .3 \rightarrow do$ sampled
HB^2: evidence $\rightarrow \neg hb \wedge w^2 = .05 \times Pr(\neg hb|do) = .015$

- a very rough estimate having 3 samples only

$$Pr(fo|lo, \neg hb) \approx \frac{.18}{.0495 + .015 + .18} = .74$$

Gibbs sampling

- sampling process:

  $\forall$ samples $p^m = \{P_1 = p_1^m, \ldots, P_n = p_n^m\}$, $m \in \{1, \ldots, M\}$

  1. fix states of all observed variables from $\mathbf{E}$ (in all samples),
  2. the other variables initialized in p^0 randomly,
  3. generate p^m from p^{m-1} $(\forall P_i \notin E)$
     - $p_1^m \leftarrow Pr(P_1|p_2^{m-1}, \ldots, p_n^{m-1})$,
     - $p_2^m \leftarrow Pr(P_2|p_1^m, p_3^{m-1}, \ldots, p_n^{m-1})$,
     - $\ldots$,
     - $p_n^m \leftarrow Pr(P_n|p_1^m, \ldots, p_{n-1}^m)$,
  4. repeat step 3 for $m \in \{1, \ldots, M\}$,

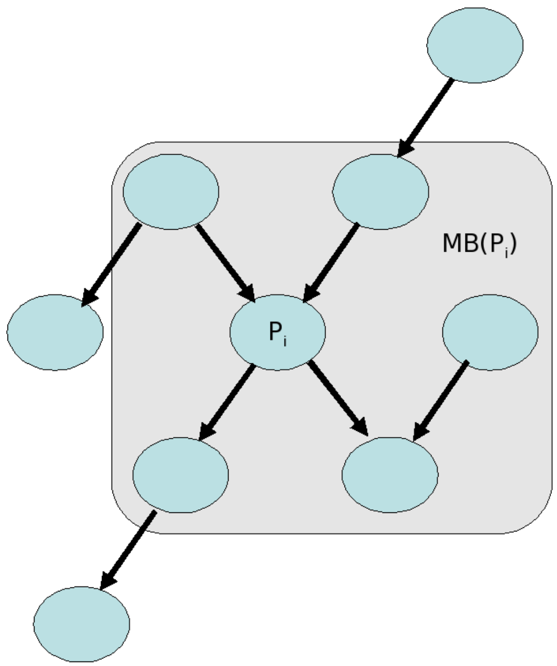
  ignore samples at the beginning (burn-in period).

Gibbs sampling

- probs $Pr(P_i|P_1, \ldots, P_{i-1}, P_{i+1}, \ldots, P_n) = Pr(P_i|P \setminus P_i)$ not explicitly given ...
  - to enumerate them, only their BN neighborhood needs to be known

$$Pr(P_i|P \setminus P_i) \propto Pr(P_i|parents(P_i)) \prod_{\forall P_j, P_i \in parents(P_j)} Pr(P_j|parents(P_j))$$

  - the neighborhood is called **Markov blanket** (MB),
  - MB covers the node, its parents, its children and their parents,
  - $MB(P_i)$ is the minimum set of nodes that d-separates P_i from the rest of the network.



- from samples to probabilities?
  - evidence holds in all samples (by definition),
  - averaging $\forall m$ is applied to find $Pr(Q|\mathbf{e})$

$$Pr(p_i|\mathbf{e}) \approx \frac{\sum_{m=1}^{M} \delta(p_i^m, p_i)}{M} \quad \delta(i,j) = \begin{cases} 1 & \text{for } i = j \\ 0 & \text{for } i \neq j \end{cases}$$

Gibbs sampling – example

- let us approximate $Pr(fo|lo, \neg hb)$ (its exact value computed earlier is 0.5),

	p^0	p^1	p^2	...
FO	T	**F**	**F**	
BP	T	**F**	**F**	
LO	**T**	**T**	**T**	
DO	F	F	F	
HB	**F**	**F**	**F**	

p^0: random init of unevidenced variables

FO^1: $Pr^*(fo) \propto Pr(fo) \times Pr(lo|fo) \times Pr(\neg do|fo, bp)$

$Pr^*(\neg fo) \propto Pr(\neg fo) \times Pr(lo|\neg fo) \times Pr(\neg do|\neg fo, bp)$

$Pr^*(fo) \propto .15 \times .6 \times .01 = 9 \times 10^{-4} \rightarrow \times\alpha^1_{FO} = .41$

$Pr^*(\neg fo) \propto .85 \times .05 \times .03 = 1.275 \times 10^{-3} \rightarrow \times\alpha^1_{FO} = .59$

$\alpha^1_{FO} = \frac{1}{Pr^*(fo)+Pr^*(\neg fo)} = 460$

BP^1: $Pr^*(bp) \propto Pr(bp) \times Pr(\neg do|\neg fo, bp) = .01 \times .03 = .0003$

$Pr^*(\neg bp) \propto Pr(\neg bp) \times Pr(\neg do|\neg fo, \neg bp) = .99 \times .7 = 0.693$

$\alpha^1_{BP} = \frac{1}{Pr^*(bp)+Pr^*(\neg bp)} = 1.44 \rightarrow Pr^*(bp) = 4 \times 10^{-4}$

DO^1: by analogy, $|MB(DO)| = 5$

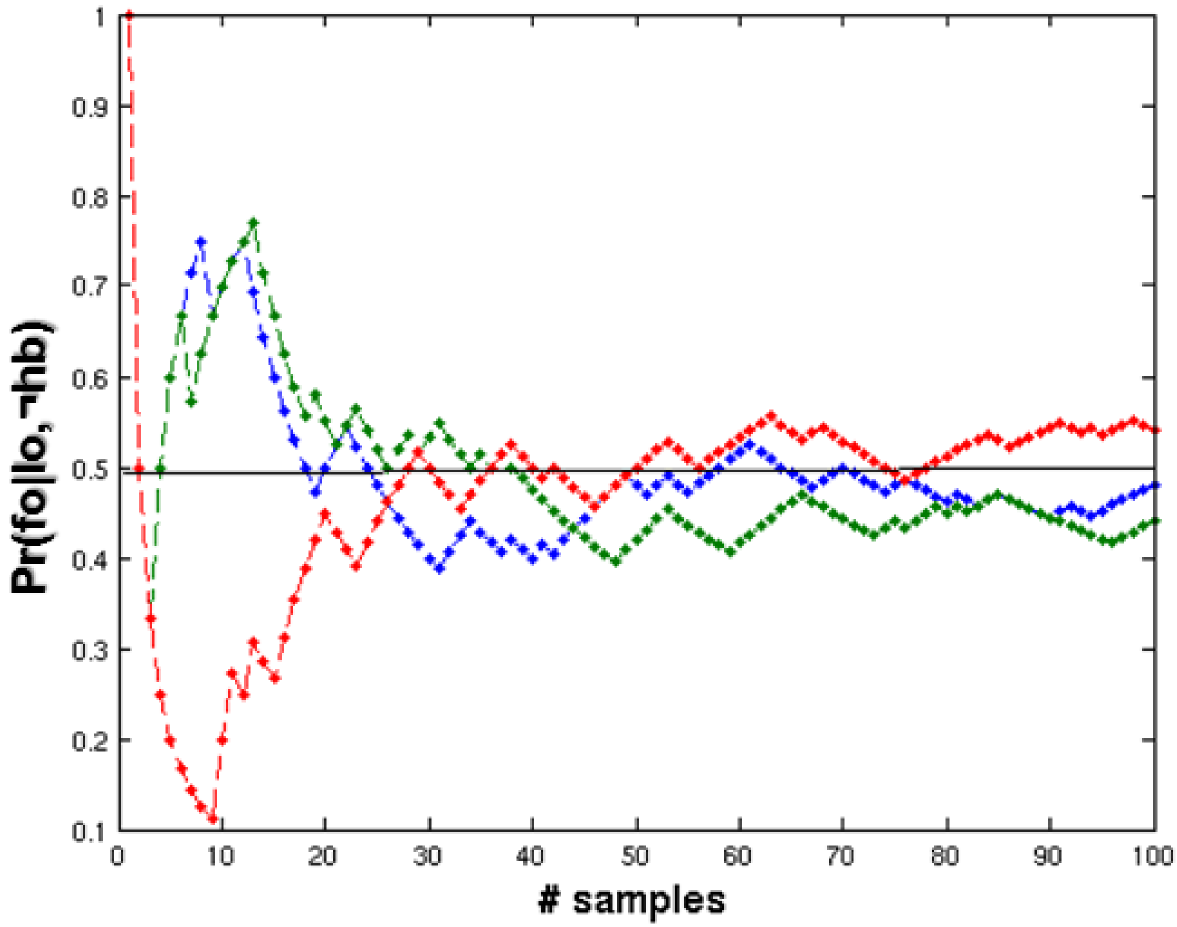
FO^2: BP value was switched, substitution is $Pr(DO|FO, \neg bp)$

$Pr^*(fo) = .21 \quad Pr^*(\neg fo) = .79$

BP^2: the same probs as is sample 1

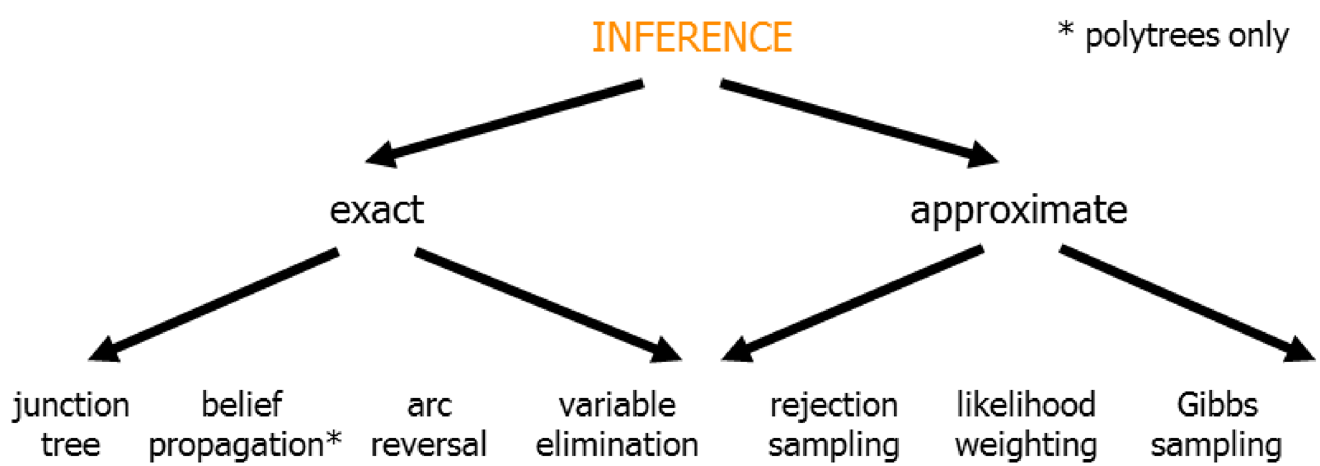
Gibbs sampling – example

- BN Matlab Toolbox, aproximation of $Pr(fo|lo, \neg hb)$,
- gibbs_sampling_inf_engine, three independent runs with 100 samples.



Summary

- independence and conditional independence remarkably simplify prob model
  - still, BN inference remains generally **NP-hard** wrt the number of nodes,
  - inference complexity grows with the number of network edges
    - naïve Bayes model – linear complexity,
    - exponential in the size of maximal clique of induced graph,
  - inference complexity can be reduced by constraining model structure
    - special network types (singly connected), e.g. trees – one parent only,
  - inference time can be shorten when exact answer not required
    - approximate inference, typically (but not only) stochastic sampling.



Recommended reading, lecture resources

- Russell, Norvig: **AI: A Modern Approach**, Uncertain Knowledge and Reasoning (Part V)
  - probabilistic reasoning (chapter 14 or 15, depends on the edition),
  - online on Google books: http://books.google.com/books?id=8jZBksh-bUMC,
  - Norvig's videos on probabilistic inference:
    - http://www.youtube.com/watch?v=q5DHnmHtVmc&feature=plcp,
- Koller, Friedman: **Probabilistic Graphical Models: Principles and Techniques.**
  - book: http://pgm.stanford.edu/, chapter II, inference, variable elimination,
  - coursera: https://www.coursera.org/course/pgm.