

CSE 473: Artificial Intelligence

Bayesian Networks: Inference

Hanna Hajishirzi

Many slides over the course adapted from either Luke Zettlemoyer, Pieter Abbeel, Dan Klein, Stuart Russell or Andrew Moore

Outline

- Bayesian Networks Inference
 - Exact Inference: Variable Elimination
 - Approximate Inference: Sampling

Approximate Inference

- Simulation has a name: sampling
- Sampling is a hot topic in machine learning, and it's really simple
- Basic idea:
 - Draw N samples from a sampling distribution S
 - Compute an approximate posterior probability
 - Show this converges to the true probability P
- Why sample?
 - Learning: get samples from a distribution you don't know
 - Inference: getting a sample is faster than computing the right answer (e.g. with variable elimination)



Sampling

- Sampling from given distribution

- Step 1: Get sample u from uniform distribution over $[0, 1)$
 - E.g. `random()` in python
- Step 2: Convert this sample u into an outcome for the given distribution by having each outcome associated with a sub-interval of $[0,1)$ with sub-interval size equal to probability of the outcome

- Example

C	P(C)
red	0.6
green	0.1
blue	0.3

$$0 \leq u < 0.6, \rightarrow C = \textit{red}$$
$$0.6 \leq u < 0.7, \rightarrow C = \textit{green}$$
$$0.7 \leq u < 1, \rightarrow C = \textit{blue}$$

- If `random()` returns $u = 0.83$, then our sample is $C = \textit{blue}$
- E.g, after sampling 8 times:

Sampling in BN

- Prior Sampling
- Rejection Sampling
- Likelihood Weighting
- Gibbs Sampling

Prior Sampling

$$P(C)$$

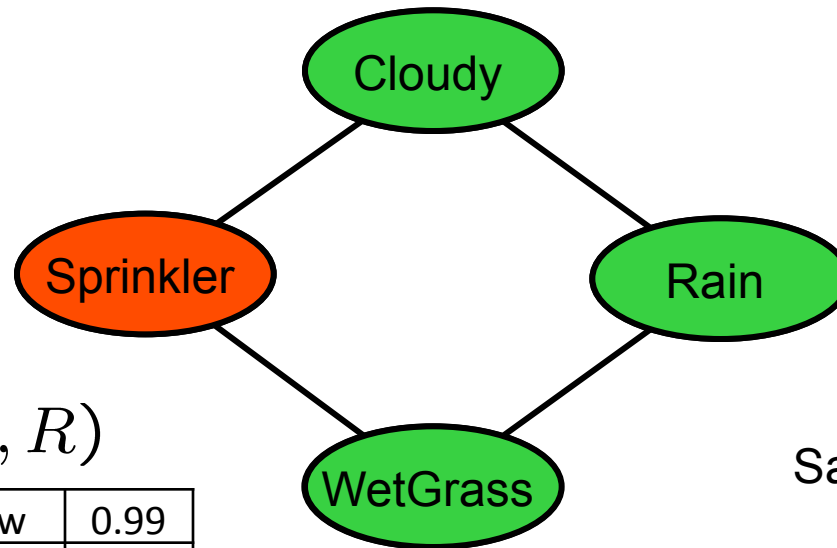
+c	0.5
-c	0.5

$$P(S|C)$$

	+s	0.1
+c	-s	0.9
	+s	0.5
-c	-s	0.5

$$P(R|C)$$

	+r	0.8
+c	-r	0.2
	+r	0.2
-c	-r	0.8



$$P(W|S, R)$$

		+w	0.99
	+r	-w	0.01
		+w	0.90
+s	-r	-w	0.10
		+w	0.90
	+r	-w	0.10
		+w	0.01
-s	-r	-w	0.99

Samples:

+c, -s, +r, +w

-c, +s, -r, +w

...

Prior Sampling

- For $i=1, 2, \dots, n$
 - Sample x_i from $P(X_i \mid \text{Parents}(X_i))$
- Return (x_1, x_2, \dots, x_n)

Prior Sampling

- This process generates samples with probability:

$$S_{PS}(x_1 \dots x_n) = \prod_{i=1}^n P(x_i | \text{Parents}(X_i)) = P(x_1 \dots x_n)$$

...i.e. the BN's joint probability

- Let the number of samples of an event be $N_{PS}(x_1 \dots x_n)$
- Then
$$\begin{aligned} \lim_{N \rightarrow \infty} \hat{P}(x_1, \dots, x_n) &= \lim_{N \rightarrow \infty} N_{PS}(x_1, \dots, x_n) / N \\ &= S_{PS}(x_1, \dots, x_n) \\ &= P(x_1 \dots x_n) \end{aligned}$$
- I.e., the sampling procedure is **consistent**

Example

- We'll get a bunch of samples from the BN:

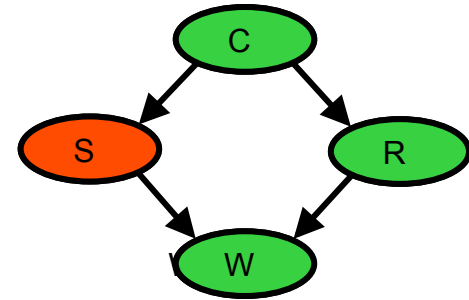
+c, -s, +r, +w

+c, +s, +r, +w

-c, +s, +r, -w

+c, -s, +r, +w

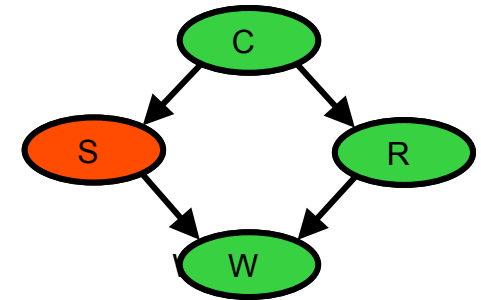
-c, -s, -r, +w



- If we want to know $P(W)$
 - We have counts $\langle +w:4, -w:1 \rangle$
 - Normalize to get $P(W) = \langle +w:0.8, -w:0.2 \rangle$
 - This will get closer to the true distribution with more samples
 - Can estimate anything else, too
 - What about $P(C | +w)$? $P(C | +r, +w)$? $P(C | -r, -w)$?
 - Fast: can use fewer samples if less time (what's the drawback?)

Rejection Sampling

- Let's say we want $P(C)$
 - No point keeping all samples around
 - Just tally counts of C as we go
- Let's say we want $P(C | +s)$
 - Same thing: tally C outcomes, but ignore (reject) samples which don't have $S=+s$
 - This is called rejection sampling
 - It is also consistent for conditional probabilities (i.e., correct in the limit)



+c, -s, +r, +w
+c, +s, +r, +w
-c, +s, +r, -w
+c, -s, +r, +w
-c, -s, -r, +w

Sampling Example

- There are 2 cups.
 - The first contains 1 penny and 1 quarter
 - The second contains 2 quarters
- Say I pick a cup uniformly at random, then pick a coin randomly from that cup. It's a quarter (yes!). What is the probability that the other coin in that cup is also a quarter?

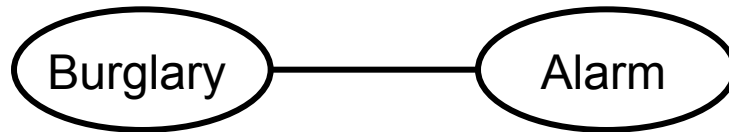
Rejection Sampling

- IN: evidence instantiation
- For $i=1, 2, \dots, n$
 - Sample x_i from $P(X_i \mid \text{Parents}(X_i))$
 - If x_i not consistent with evidence
 - Reject: Return, and no sample is generated in this cycle
- Return (x_1, x_2, \dots, x_n)

Likelihood Weighting

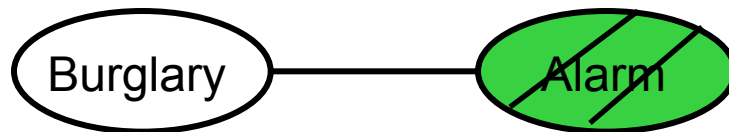
- Problem with rejection sampling:

- If evidence is unlikely, you reject a lot of samples
- You don't exploit your evidence as you sample
- Consider $P(B|+a)$



-b, -a
-b, -a
-b, -a
-b, -a
+b, +a

- Idea: fix evidence variables and sample the rest



-b +a
-b, +a
-b, +a
-b, +a
+b, +a

- Problem: sample distribution not consistent!
- Solution: weight by probability of evidence given parents

Likelihood Weighting

$$P(C)$$

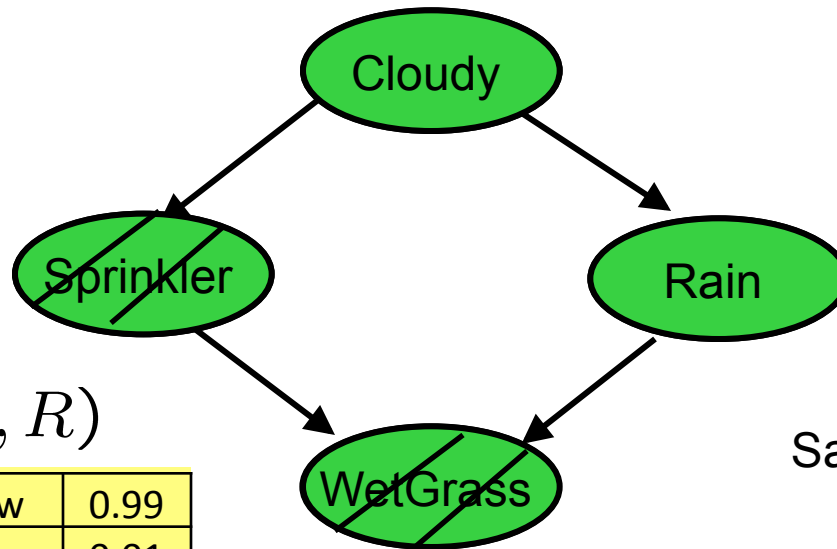
+c	0.5
-c	0.5

$$P(S|C)$$

	+s	0.1
+c	-s	0.9
	+s	0.5
-c	-s	0.5

$$P(R|C)$$

	+r	0.8
+c	-r	0.2
	+r	0.2
-c	-r	0.8



$$P(W|S, R)$$

		+w	0.99
	+r	-w	0.01
+s	-r	+w	0.90
		-w	0.10
-s	+r	+w	0.90
		-w	0.10
	-r	+w	0.01
		-w	0.99

Samples:

+c, +s, +r, +w

...

$$w = 1.0 \times 0.1 \times 0.99$$

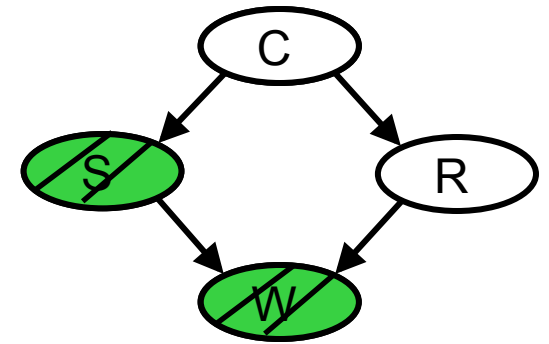
Likelihood Weighting

- Sampling distribution if z sampled and e fixed evidence

$$S_{WS}(z, e) = \prod_{i=1}^l P(z_i | \text{Parents}(Z_i))$$

- Now, samples have weights

$$w(z, e) = \prod_{i=1}^m P(e_i | \text{Parents}(E_i))$$



- Together, weighted sampling distribution is consistent

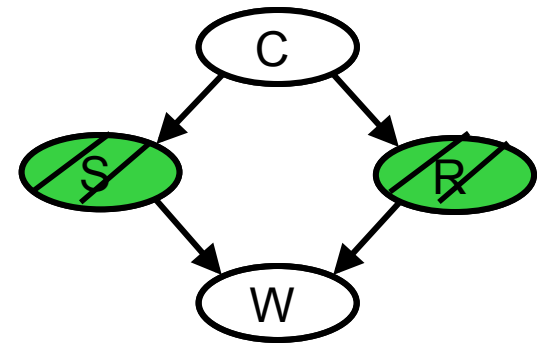
$$\begin{aligned} S_{WS}(z, e) \cdot w(z, e) &= \prod_{i=1}^l P(z_i | \text{Parents}(z_i)) \prod_{i=1}^m P(e_i | \text{Parents}(e_i)) \\ &= P(z, e) \end{aligned}$$

Likelihood Weighting

- IN: evidence instantiation
- $w = 1.0$
- for $i=1, 2, \dots, n$
 - if X_i is an evidence variable
 - $X_i =$ observation x_i for X_i
 - Set $w = w * P(x_i | \text{Parents}(X_i))$
 - else
 - Sample x_i from $P(X_i | \text{Parents}(X_i))$
- return $(x_1, x_2, \dots, x_n), w$

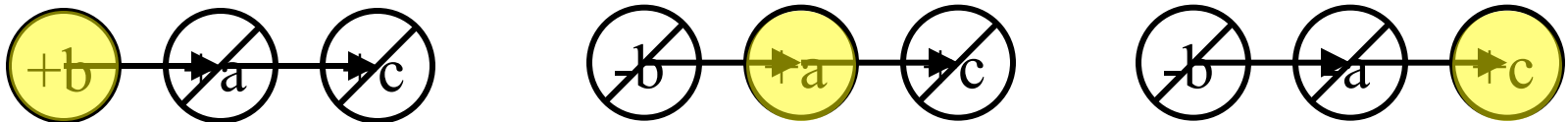
Likelihood Weighting

- Likelihood weighting is good
 - We have taken evidence into account as we generate the sample
 - E.g. here, W 's value will get picked based on the evidence values of S , R
 - More of our samples will reflect the state of the world suggested by the evidence
- Likelihood weighting doesn't solve all our problems
 - Evidence influences the choice of downstream variables, but not upstream ones (C isn't more likely to get a value matching the evidence)
- We would like to consider evidence when we sample every variable



Markov Chain Monte Carlo*

- Idea: instead of sampling from scratch, create samples that are each like the last one.
- Gibbs Sampling: resample one variable at a time, conditioned on the rest, but keep evidence fixed.



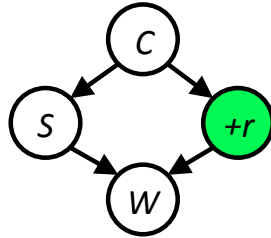
- Properties: Now samples are not independent (in fact they're nearly identical), but sample averages are still consistent estimators!
- What's the point: both upstream and downstream variables condition on evidence.

Gibbs Sampling Example

$P(S|+r)$

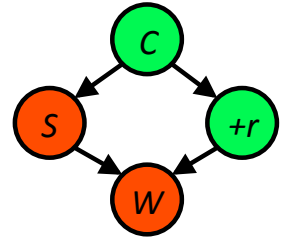
- Step 1: Fix evidence

- $R = +r$



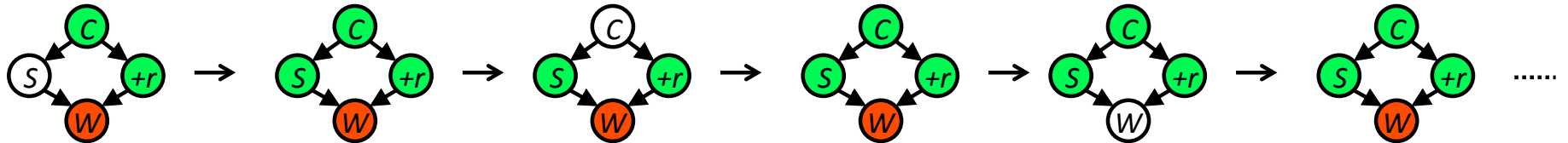
- Step 2: Initialize other variables

- Randomly



- Steps 3: Repeat

- Choose a non-evidence variable X
- Resample X from $P(X | \text{all other variables})$



Sample from $P(S | +c, -w, +r)$

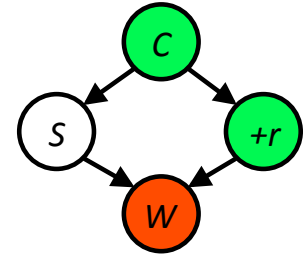
Sample from $P(C | +s, -w, +r)$

Sample from $P(W | +s, +c, +r)$

Sampling One Variable

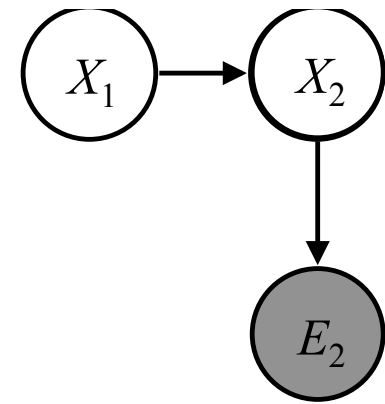
- Sample from $P(S \mid +c, +r, -w)$

$$\begin{aligned} P(S \mid +c, +r, -w) &= \frac{P(S, +c, +r, -w)}{P(+c, +r, -w)} \\ &= \frac{P(S, +c, +r, -w)}{\sum_s P(s, +c, +r, -w)} \\ &= \frac{P(+c)P(S \mid +c)P(+r \mid +c)P(-w \mid S, +r)}{\sum_s P(+c)P(s \mid +c)P(+r \mid +c)P(-w \mid s, +r)} \\ &= \frac{P(+c)P(S \mid +c)P(+r \mid +c)P(-w \mid S, +r)}{P(+c)P(+r \mid +c) \sum_s P(s \mid +c)P(-w \mid s, +r)} \\ &= \frac{P(S \mid +c)P(-w \mid S, +r)}{\sum_s P(s \mid +c)P(-w \mid s, +r)} \end{aligned}$$

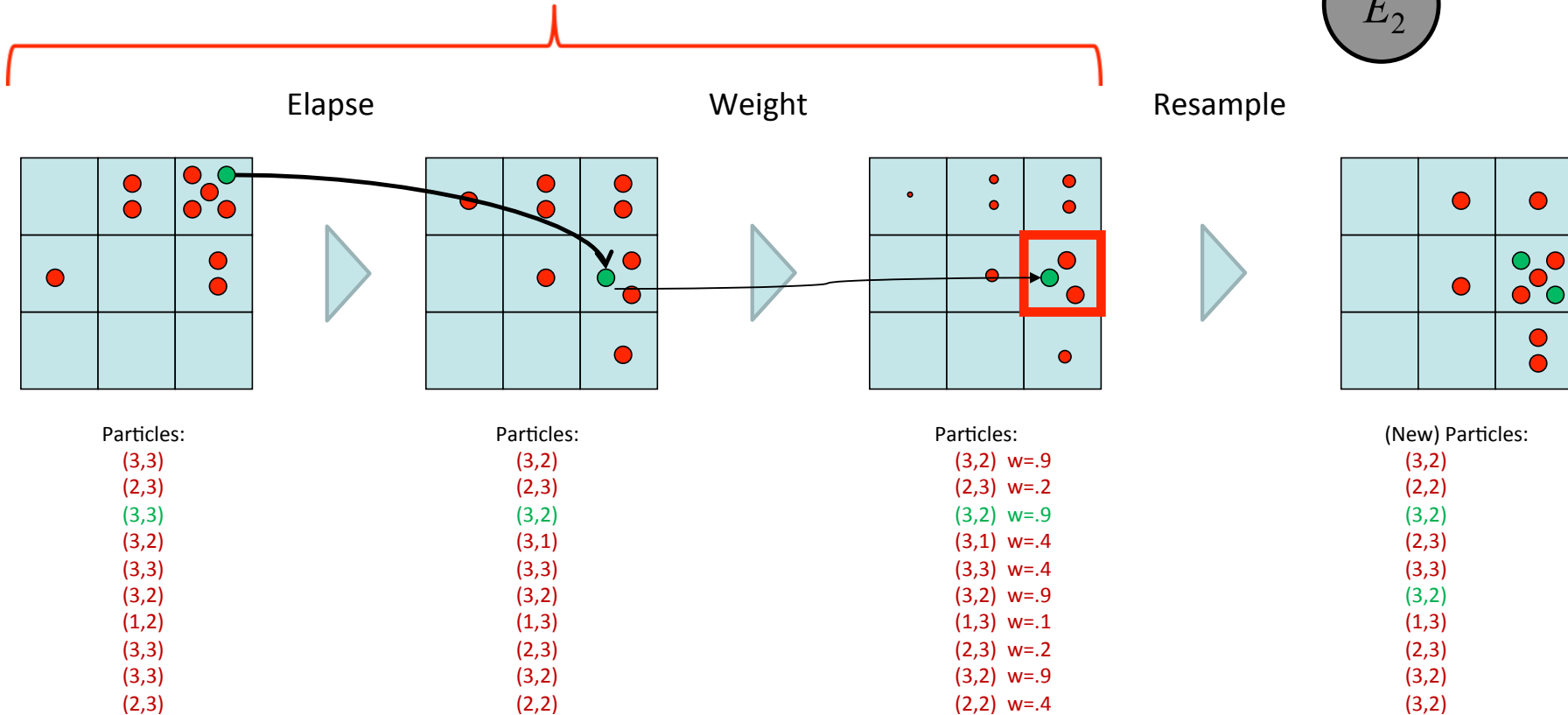


- Many things cancel out – only CPTs with S remain!
- More generally: only CPTs that have resampled variable need to be considered, and joined together

How About Particle Filtering?

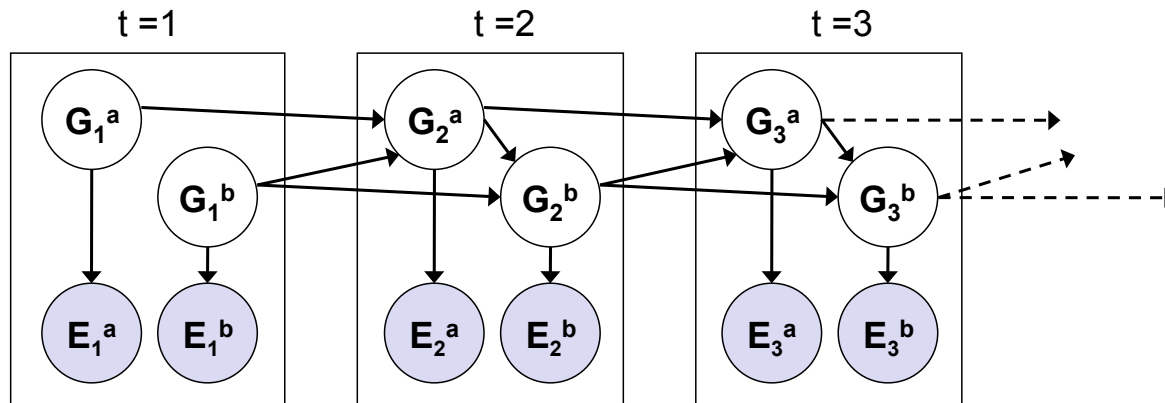


= likelihood weighting



Dynamic Bayes Nets (DBNs)

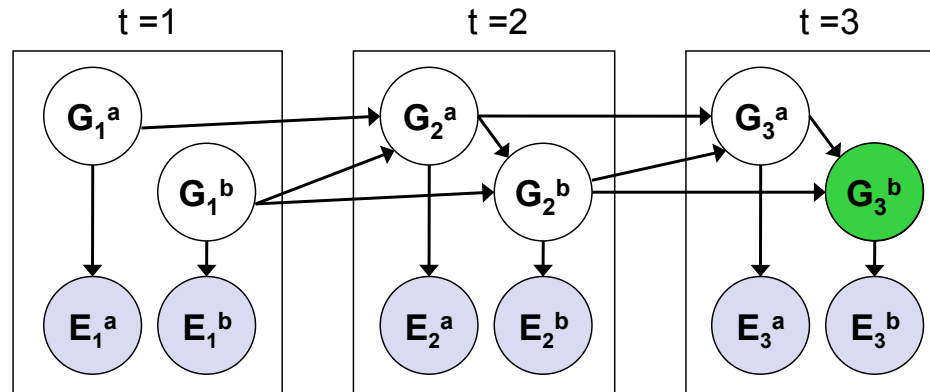
- We want to track multiple variables over time, using multiple sources of evidence
- Idea: Repeat a fixed Bayes net structure at each time
- Variables from time t can condition on those from $t-1$



- Discrete valued dynamic Bayes nets (with evidence on the bottom) are HMMs

Exact Inference in DBNs

- Variable elimination applies to dynamic Bayes nets
- Procedure: “unroll” the network for T time steps, then eliminate variables until $P(X_T | e_{1:T})$ is computed



- Online belief updates: Eliminate all variables from the previous time step; store factors for current time only

Particle Filtering in DBNs

- A particle is a complete sample for a time step
- **Initialize:** Generate prior samples for the $t=1$ Bayes net
 - Example particle: $\mathbf{G}_1^a = (3,3)$ $\mathbf{G}_1^b = (5,3)$
- **Elapse time:** Sample a successor for each particle
 - Example successor: $\mathbf{G}_2^a = (2,3)$ $\mathbf{G}_2^b = (6,3)$
- **Observe:** Weight each entire sample by the likelihood of the evidence conditioned on the sample
 - Likelihood: $P(\mathbf{E}_1^a | \mathbf{G}_1^a) * P(\mathbf{E}_1^b | \mathbf{G}_1^b)$
- **Resample:** Select prior samples (tuples of values) in proportion to their likelihood