CSC384: Lecture 9

- Last time
  - STRIPS and Regression planning
- Today
  - Reasoning under uncertainty (intro)
- Readings:
  - Today: 10.1 – 10.3 (reasoning under uncertainty)
  - Next week: 10.3 (belief networks)

Reasoning Under Uncertainty

- So far our planning problems have assumed:
  - start state is *known with certainty*
  - actions are *deterministic*
  - Both assumptions unrealistic (e.g., in robot domain)
- Knowledge:
  - What if Craig can be in office *or* lab?
  - Can robot know a priori if coffee made? mail waits?
- Actions:
  - Robbie grabs coffee: could fail (try again); could spill (make more)
  - Robbie does move(l,o): could end up elsewhere

Classical Plans are Inadequate

- Classical plans (sequences of actions) fail to handle either of these
  - similarly for “deterministic domain” search procedures
- If actions have nondeterministic (uncertain) effects, suitable plan should *branch* on outcome
  - Pour coffee; **If** successful, go to office, **else** make more
- If uncertain knowledge, uncertain effects, there may be no plan guaranteed to achieve goal
  - Pour coffee; If spilled, there is no more and plan fails

Some Decision Making Scenarios

- Suppose Craig wants coffee (as quickly as possible). Robbie, with coffee ready to deliver, knows Craig is in lab ***or*** office, but not which. Robbie is much closer to the lab than the office. What should it do?
  - What if Craig is *almost* always in his office?
  - What if peeking into the lab takes just a couple seconds? a couple minutes?
  - What if Craig is equally likely to be lab or office?

More Decision Making Scenarios

- Suppose Craig wants coffee or his mail (doesn't need both). Robbie only as enough battery power to do one or the other. What goal should it (attempt) to achieve?
  - What if Craig likes coffee much more than mail? mail more than coffee?
  - What if possibility no coffee made?
  - What if Robbie usually spills coffee a couple of times before succeeding?
- Aside: consider the lottery. What would have to change for you to start (or stop) playing?

Degrees of Belief and Preference

- The *right* decision/plan above requires that Robbie consider how important various objectives are, how likely he is to achieve them, and make tradeoffs among them.
- This generally requires that we *quantify* our preferences and our degrees of belief.
- We're going to start with *degrees of belief* and talk about preferences and decision making later.

Reasoning under Uncertainty

- Logical representations (as seen so far) capture a weak form of uncertainty:
  - if query *q* not proved, could be true or false
- Too weak to allow (reasonable) tradeoffs
  - want to know *how likely* *q* is
- We'll *quantify* our beliefs using *probabilities*
  - *Pr(q)* denotes probability that you believe *q* is true
  - We take subjectivist viewpoint (cf. *frequentist*)
- Note: statistics/data *influence* degrees of belief
- Let's formalize things…

Random Variables

- Assume set ***V*** of *random variables*: *X*, *Y*, etc.
  - Each RV *X* has a *domain* of values *Dom(X)*
  - *X* can take on any value from *Dom(X)*
  - Assume **V** and *Dom(X)* finite
- Examples
  - *Dom(X)* = $\{x_1, x_2, x_3\}$
  - *Dom(Weather) = {sunny, cloudy, rainy}*
  - *Dom(StudentInCraigsOffice) =*

    {pascal, georgios, veronica, tianhan…}
  - *Dom(CraigHasCoffee) = {T,F}* (boolean var)

Random Variables/Possible Worlds

- A *formula* is a logical combination of variable assignments:
  - $X = x_1$; $(X = x_2 \vee X = x_3) \wedge Y = y_2$; $(x_2 \vee x_3) \wedge y_2$
  - chc ∧ ~cm, etc…
  - let $\mathcal{L}$ denote the set of formulae (our language)
- A *possible world* is an assignment of values to each RV
  - these are analogous to truth assts (interpretations)
  - Let W be the set of worlds

Probability Distributions

- A probability distribution Pr: $\mathcal{L} \rightarrow [0,1]$ s.t.
  - $0 \leq Pr(\alpha) \leq 1$
  - $Pr(\alpha) = Pr(\beta)$ if α is logically equivalent to β
  - $Pr(\alpha) = 1$ if α is a tautology
  - $Pr(\alpha \vee \beta) = Pr(\alpha) + Pr(\beta) - Pr(\alpha \wedge \beta)$
- $Pr(\alpha)$ denotes our *degree of belief* in α; e.g.
  - $Pr(X = x_1) = Pr(x_1) = 0.9$
  - $Pr((x_2 \vee x_3) \wedge y_2) = 0.9$
  - Pr(loc(craig) = off) = 0.6
  - Pr(loc(craig) = off ∨ loc(craig) = lab) = 1.0
  - Pr(loc(craig) = lounge) = 0.0

Semantics of Prob. Distributions

- A probability measure μ: W → [0,1] s.t.
  - $$\sum_{w \in W} \mu(w) = 1$$

- Intuitively, μ(w) measures the probability that the actual world is w (your *belief* in w). Thus, the relative likelihood of any world you consider possible is specified. If w has measure 0, you consider it to be impossible!

Semantics of Distributions

- Given measure μ, we determine degree of belief in formula Pr(α)
  - simply sum the measures of all worlds satisfying the formula of interest

$$\Pr(\alpha) = \sum_{w \in W} \{\mu(w) : w \models \alpha\}$$

Example Distribution

T – mail truck outside
M – mail waiting
C – craig wants coffee
A – craig is angry

$t\ c\ m\ a$	0.162	$\bar{t}\ c\ m\ a$	0.0
$t\ c\ m\ \bar{a}$	0.018	$\bar{t}\ c\ m\ \bar{a}$	0.0
$t\ c\ \bar{m}\ a$	0.016	$\bar{t}\ c\ \bar{m}\ a$	0.0
$t\ c\ \bar{m}\ \bar{a}$	0.004	$\bar{t}\ c\ \bar{m}\ \bar{a}$	0.0
$t\ \bar{c}\ m\ a$	0.432	$\bar{t}\ \bar{c}\ m\ a$	0.0
$t\ \bar{c}\ m\ \bar{a}$	0.288	$\bar{t}\ \bar{c}\ m\ \bar{a}$	0.0
$t\ \bar{c}\ \bar{m}\ a$	0.008	$\bar{t}\ \bar{c}\ \bar{m}\ a$	0.0
$t\ \bar{c}\ \bar{m}\ \bar{a}$	0.072	$\bar{t}\ \bar{c}\ \bar{m}\ \bar{a}$	0.0

Pr(t) =1

Pr(-t) = 0

Pr(c) = .2

Pr(-c) = .8

Pr(m) = .9

Pr(a) = .618

Pr(c & m) = .18

Pr(c v m) = .92

Pr(a -> m)

= Pr(-a v m)

= 1 – Pr(a & -m)

= .976

Relationship

- For any measure μ the induced mapping Pr is a distribution.
- For any distribution Pr there is a corresponding μ measure that induces Pr.
- Thus, the syntactic and semantic restrictions correspond (soundness and completeness)

Some Important Properties

- Pr(α) = 1 – Pr(-α), where α can be "generalized"
- $$\sum \{\Pr(x) : x \in Dom(X)\} = 1$$
  - e.g., Pr(sunny) + Pr(cloudy) + Pr(rainy) = 1
- Pr(αvβ) =1 if α ⊢ -β
- $$\Pr(x) = \sum_{y \in Dom(Y)} \Pr(x \wedge y)$$
  - this is called the *summing out property*
  - e.g., Pr(a) = Pr(a & m) + Pr(a & -m)

Conditional Probability

- Conditional probability critical in inference

$$\Pr(b \mid a) = \frac{\Pr(b \wedge a)}{\Pr(a)}$$

  - if Pr(a) = 0, we often treat Pr(b|a)=1 by convention

Semantics of Conditional Prob.

- Semantics of Pr(b|a):
  - denotes relative weight of b-worlds among a-worlds
  - ~a-worlds play no role

$$\Pr(b \mid a) = \frac{\sum \{\mu(w) : w \models a \wedge b\}}{\sum \{\mu(w) : w \models a\}}$$

Intuitive Meaning of Cond. Prob.

- Intuitively, if you learned a, you would change your degree of belief in b from Pr(b) to Pr(b|a)
- In our example:
  - Pr(m|c) = 0.9
  - Pr(m| ~c) = 0.9
  - Pr(a) = 0.618
  - Pr(a|~m) = 0.27
  - Pr(a|~m & c) = 0.8
- Notice the *nonmonotonicity* in the last three cases when additional evidence is added
  - contrast this with logical inference

Some Important Properties

- **Product Rule:** $\Pr(ab) = \Pr(a|b)\Pr(b)$

- **Summing Out Rule:**

$$\Pr(a) = \sum_{b \in Dom(B)} \Pr(a \mid b)\Pr(b)$$

- **Chain Rule:**

$\Pr(abcd) = \Pr(a|bcd)\Pr(b|cd)\Pr(c|d)\Pr(d)$

  - holds for any number of variables

Bayes Rule

- **Bayes Rule:**

$$\Pr(a \mid b) = \frac{\Pr(b \mid a)\Pr(a)}{\Pr(b)}$$

- Bayes rule follows by simple algebraic manipulation of the defn of condition probability
  - why is it so important? why significant?
  - usually, one “direction” easier to assess than other

Example of Use of Bayes Rule

- Disease ∈ {malaria, cold, flu}; Symptom = fever
  - Must compute Pr(D|fever) to prescribe treatment
- Why not assess this quantity directly?
  - Pr(mal | fever) is not natural to assess; Pr(fever | mal) reflects the underlying "causal" mechanism
  - Pr(mal | fever) is not "stable": a malaria epidemic changes this quantity (for example)
- So we use Bayes rule:
  - Pr(mal | fever) = Pr(fever | mal) Pr(mal) / Pr(fever)
  - note that Pr(fev) = Pr(m | fev) + Pr(c|fev) + Pr(fl | fev)
  - so if we compute Pr of each disease given fever using Bayes rule, normalizing constant is "free"

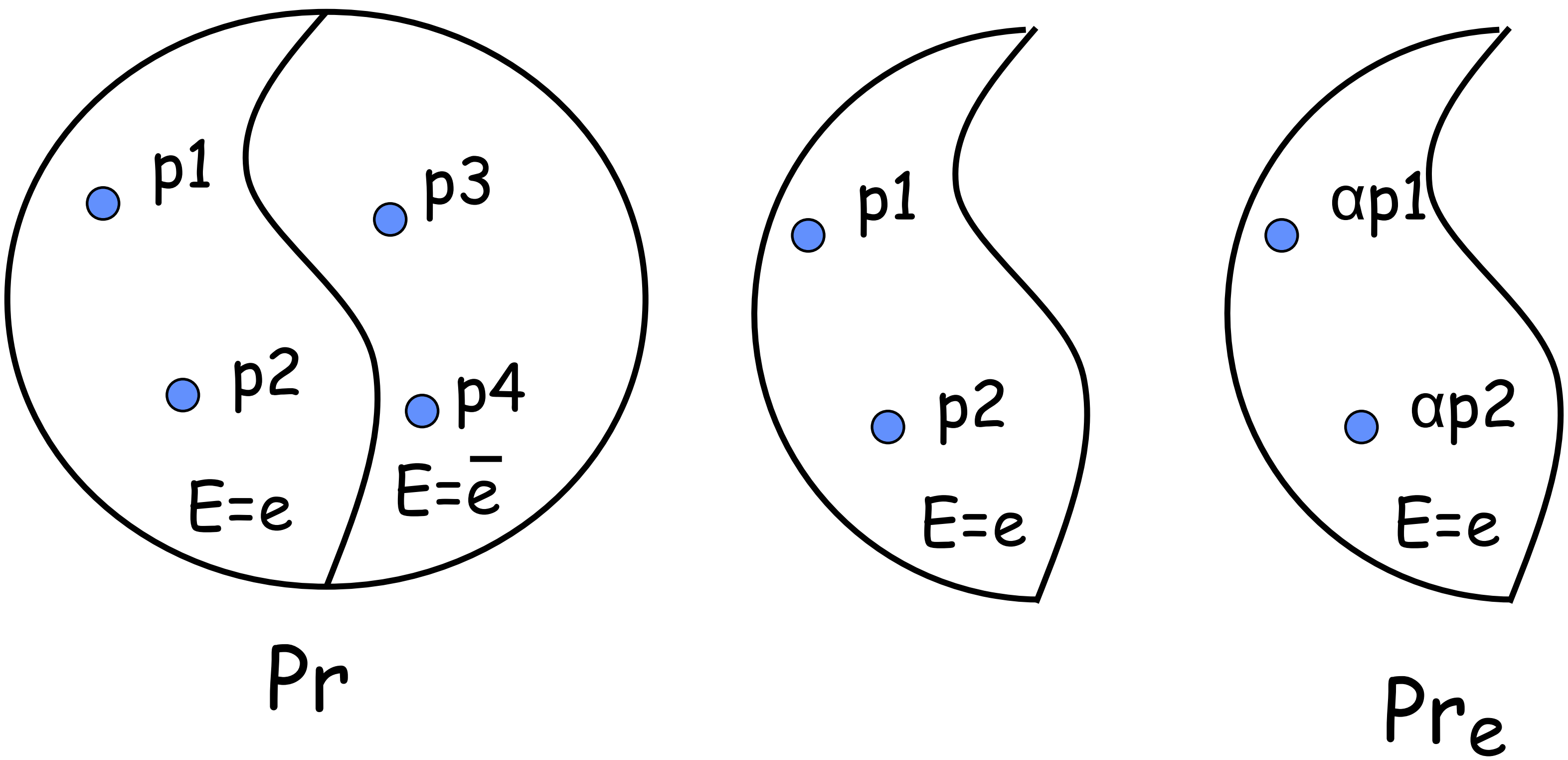
Probabilistic Inference

- By probabilistic inference, we mean
  - given a *prior* distribution *Pr* over variables of interest, representing degrees of belief
  - and given new evidence $E=e$ for some var E
  - Revise your degrees of belief: *posterior* Pr_e
- How do your degrees of belief change as a result of learning $E=e$ (or more generally $\mathbf{E}=\mathbf{e}$, for set $\mathbf{E}$)
- Contrast with "logical" reasoning over probabilistic assertions

Conditioning

- We define $Pr_e(\alpha) = Pr(\alpha / e)$
- That is, we produce Pr_e by *conditioning* the prior distribution on the observed evidence e
- Semantically, we take original measure μ
  - we set μ(w) = 0 for any world falsifying e
  - we set μ(w) = μ(w) / Pr(e) for any e-world
  - last step known as normalization (ensures that the new measure sums to 1)

Semantics of Conditioning



$\alpha = 1/(p1+p2)$
normalizing constant

Inference: Computational Bottleneck

- Semantically/conceptually, picture is clear; but several issues must be addressed
- Issue 1: How do we specify the full joint distribution over $X_1, X_2, \ldots, X_n$?
  - exponential number of possible worlds
  - e.g., if the X_i are boolean, then 2^n numbers (or 2^n -1 parameters/degrees of freedom, since they sum to 1)
  - these numbers are not robust/stable
  - these numbers are not natural to assess (what is probability that "Craig wants coffee; it's raining in Coquitlam; robot charge level is low; …"?)

Inference: Computational Bottleneck

- Issue 2: Inference in this rep'n frightfully slow
  - Must sum over exponential number of worlds to answer query $Pr(\alpha)$ or to condition on evidence e to determine $Pr_e(\alpha)$
- How do we avoid these two problems?
  - no solution in general
  - but in practice there is structure we can exploit
- We'll use conditional independence

Independence

- Recall that x and y are *independent* iff:
  - Pr(x) = Pr(x|y) iff Pr(y) = Pr(y|x) iff Pr(xy) = Pr(x)Pr(y)
  - intuitively, learning y doesn't influence beliefs about x
- x and y are *conditionally independent given* z iff:
  - Pr(x|z) = Pr(x|yz) iff Pr(y|z) = Pr(y|xz) iff
    Pr(xy|z) = Pr(x|z)Pr(y|z) iff …
  - intuitively, learning y doesn't influence your beliefs about x *if you already know z*
  - e.g., learning someone's mark on 384 exam can influence the probability you assign to a specific GPA; but if you already knew **final** 384 grade, learning the exam mark would *not* influence your GPA assessmnt

Variable Independence

- Two *variables* X and Y are conditionally independent given variable Z iff x, y are conditionally independent given z for all $x \in Dom(X)$, $y \in Dom(Y)$, $z \in Dom(Z)$
  - Also applies to sets of variables **X, Y, Z**
  - Also to unconditional case (X,Y independent)
- If you know the value of Z (*whatever* it is), nothing you learn about Y will influence your beliefs about X
  - these defns differ from earlier ones (which talk about events, not variables)

What does independence buys us?

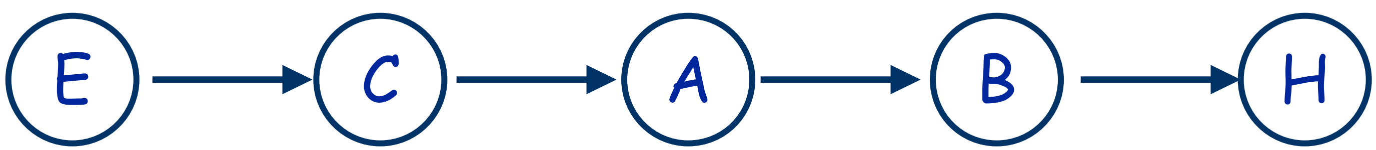
- Suppose (say, boolean) variables $X_1, X_2, \ldots, X_n$ are mutually independent
  - we can specify full joint distribution using only n parameters (linear) instead of $2^n -1$ (exponential)
- How? Simply specify $Pr(x_1), \ldots Pr(x_n)$
  - from this I can recover probability of any world or any (conjunctive) query easily
  - e.g. $Pr(x_1 \sim x_2 x_3 x_4) = Pr(x_1)\ (1-Pr(x_2))\ Pr(x_3)\ Pr(x_4)$
  - we can condition on observed value $X_k = x_k$ trivially by changing $Pr(x_k)$ to 1, leaving $Pr(x_i)$ untouched for $i \neq k$

The Value of Independence

- Complete independence reduces both *representation of joint* and *inference* from $O(2^n)$ to $O(n)$: pretty significant!
- Unfortunately, such complete mutual independence is very rare. Most realistic domains do not exhibit this property.
- Fortunately, most domains do exhibit a fair amount of conditional independence. And we can exploit conditional independence for representation and inference as well.
- **Bayesian networks** do just this

Exploiting Conditional Independence

- Let's see what conditional independence buys us
- Consider a story:
  - If Craig woke up too early E, Craig probably needs coffee C; if C, Craig needs coffee, he's likely angry A. If A, there is an increased chance of an aneurysm (burst blood vessel) B. If B, Craig is quite likely to be hospitalized H.



E – Craig woke too early A – Craig is angry H – Craig hospitalized
C – Craig needs coffee B – Craig burst a blood vessel

An Aside on Notation

- Pr(X) for variable X (or set of variables) refers to the *(marginal) distribution* over X. Pr(X|Y) refers to family of conditional distributions over X, one for each y∈Dom(Y).
- Distinguish between Pr(X) -- which is a distribution – and Pr(x) or Pr(~x) (or Pr(x_i) for nonboolean vars) -- which are numbers. Think of Pr(X) as a function that accepts any x_i ∈Dom(X) as an argument and returns Pr(x_i).
- Similarly, think of Pr(X|Y) as a function that accepts any x_i and y_k and returns Pr(x_i | y_k). Note that Pr(X|Y) is not a single distribution; rather it denotes the family of distributions (over X) induced by the different y_k ∈Dom(Y)

Cond'l Independence in our Story

- If you learned any of E, C, A, or B, your assessment of Pr(H) would change.
  - E.g., if any of these are seen to be true, you would increase Pr(h) and decrease Pr(~h).
  - So H is *not independent* of E, or C, or A, or B.
- But if you knew value of B (true or false), learning value of E, C, or A, would not influence Pr(H). Influence these factors have on H is mediated by their influence on B.
  - Craig doesn't get sent to the hospital because he's angry, he gets sent because he's had an aneurysm.
  - So H is *independent* of E, and C, and A, *given* B

Cond'l Independence in our Story

- So H is *independent* of E, and C, and A, *given* B
- Similarly:
  - B is *independent* of E, and C, *given* A
  - A is *independent* of E, *given* C
- This means that:
  - Pr(H | B, {A,C,E}) = Pr(H|B)
    - i.e., for any subset of {A,C,E}, this relation holds
  - Pr(B | A, {C,E}) = Pr(B | A)
  - Pr(A | C, {E}) = Pr(A | C)
  - Pr(C | E) and Pr(E) don't "simplify"

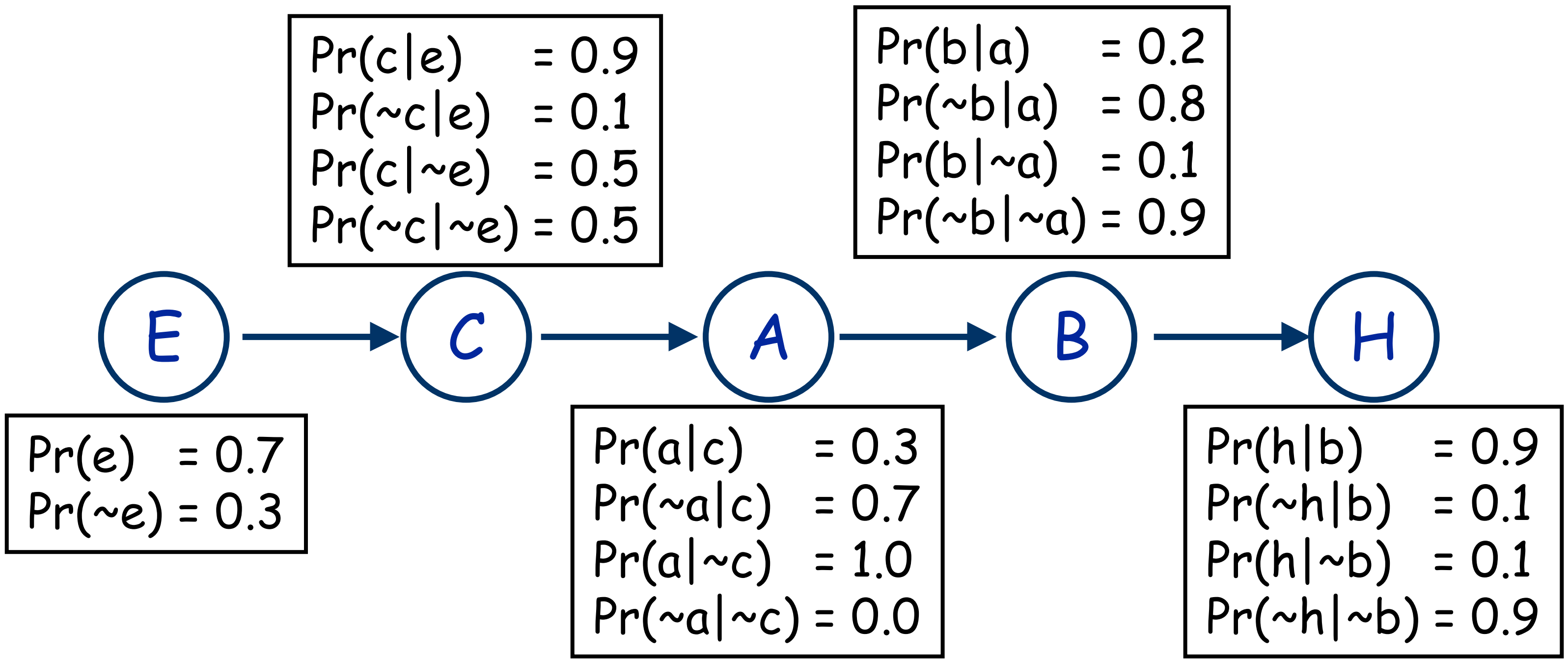
Cond'l Independence in our Story

- By the chain rule (for any instantiation of H…E):
  - Pr(H,B,A,C,E) =

    Pr(H|B,A,C,E) Pr(B|A,C,E) Pr(A|C,E) Pr(C|E) Pr(E)
- By our independence assumptions:
  - Pr(H,B,A,C,E) =

    Pr(H|B) Pr(B|A) Pr(A|C) Pr(C|E) Pr(E)
- We can specify the full joint by specifying five *local conditional distributions*: Pr(H|B); Pr(B|A); Pr(A|C); Pr(C|E); and Pr(E)

Example Quantification



- Specifying the joint requires only 9 parameters (if we note that half of these are “1 minus” the others), instead of 31 for explicit representation
  - linear in number of vars instead of exponential!
  - linear generally if dependence has a chain structure

Inference is Easy

- Want to know P(a)? Use summing out rule:

$$P(a) = \sum_{c_i \in Dom(C)} \Pr(a \mid c_i)\Pr(c_i)$$

$$= \sum_{c_i \in Dom(C)} \Pr(a \mid c_i) \sum_{e_i \in Dom(E)} \Pr(c_i \mid e_i)\Pr(e_i)$$

These are all terms specified in our local distributions!

Inference is Easy

- Computing P(a) in more concrete terms:
  - P(c) = P(c|e)P(e) + P(c|~e)P(~e)

    = 0.8 * 0.7 + 0.5 * 0.3 = 0.78
  - P(~c) = P(~c|e)P(e) + P(~c|~e)P(~e) = 0.22
    - P(~c) = 1 – P(c), as well
  - P(a) = P(a|c)P(c) + P(a|~c)P(~c)

    = 0.7 * 0.78 + 0.0 * 0.22 = 0.546
  - P(~a) = 1 – P(a) = 0.454

Bayesian Networks

- The structure above is a *Bayesian network*. A BN is a *graphical representation* of the direct dependencies over a set of variables, together with a set of *conditional probability tables* quantifying the strength of those influences.
- Bayes nets generalize the above ideas in very interesting ways, leading to effective means of representation and inference under uncertainty.
- We'll get into the details and formal definitions next time!