Adapted from AIMA slides

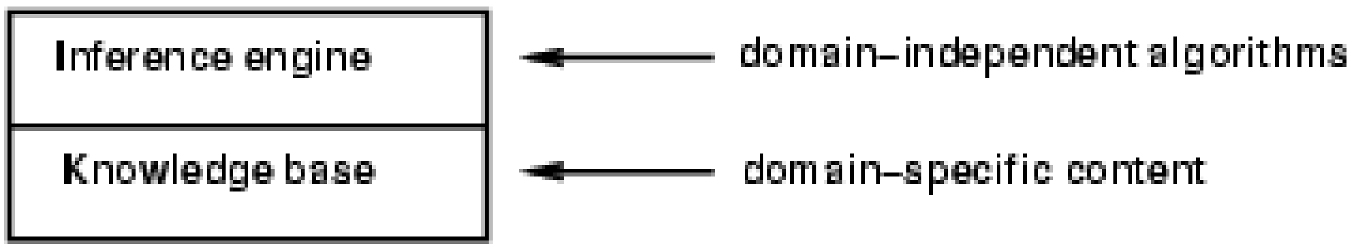
Uncertainty

Peter Antal
antal@mit.bme.hu

Outline

- Reminder
- A real-life example & demo for the homework
- Uncertainty
- Probability
- Syntax and Semantics
- Inference
- Independence and Bayes' Rule

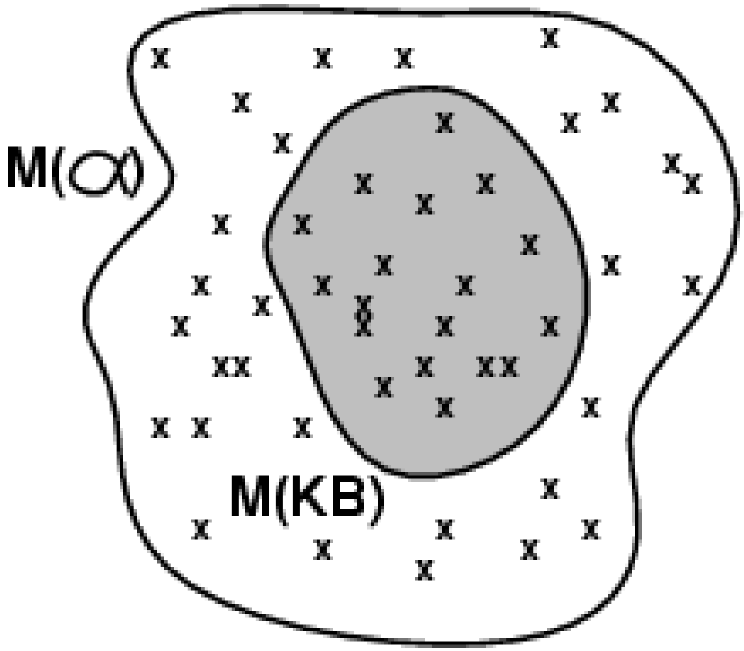
Reminder: Knowledge bases



- Knowledge base = set of sentences in a formal language
- Declarative approach to building an agent (or other system):
  - Tell it what it needs to know
  -
- Then it can Ask itself what to do - answers should follow from the KB
-
- Agents can be viewed at the knowledge level
  i.e., what they know, regardless of how implemented
- Or at the implementation level
  - i.e., data structures in KB and algorithms that manipulate them
  -

Reminder: Models

- Logicians typically think in terms of models, which are formally structured worlds with respect to which truth can be evaluated
-
- We say *m* is a model of a sentence
- *M(α)* is the set of all models of α
-
- Then KB $\models$ α iff $M(KB) \subseteq M(\alpha)$
-
  - E.g. *KB* = Giants won and Reds won α = Giants won
  -



Reminder: truth vs. proof

- Soundness: *i* is sound if whenever $KB \vdash_i \alpha$, it is also true that $KB \models \alpha$
- Completeness: *i* is complete if whenever $KB \models \alpha$, it is also true that $KB \vdash_i \alpha$
- Preview: we will define a logic (first-order logic) which is expressive enough to say almost anything of interest, and for which there exists a sound and complete inference procedure.
- That is, the procedure will answer any question whose answer follows from what is known by the *KB*.

Classical propositional logic models and inference

P_1	…	P_3	KB	s
F	F	F	F	T
F	F	T	T	F
F	T	F	F	T
F	T	T	F	F
T	F	F	F	T
T	F	T	T	T
T	T	F	F	T
T	T	T	F	T

Uncertainty

Let action A_t = leave for airport $_t$ minutes before flight
Will A_t get me there on time?

Problems:

1. partial observability (road state, other drivers' plans, etc.)
2. noisy sensors (traffic reports)
3. uncertainty in action outcomes (flat tire, etc.)
4. immense complexity of modeling and predicting traffic

Hence a purely logical approach either

1. risks falsehood: "A_{25} will get me there on time", or
2. leads to conclusions that are too weak for decision making:

"A_{25} will get me there on time if there's no accident on the bridge and it doesn't rain and my tires remain intact etc etc."

(A_{1440} might reasonably be said to get me there on time but I'd have to stay overnight in the airport …)

Methods for handling uncertainty

- Default or nonmonotonic logic:
  - Assume my car does not have a flat tire
  - Assume A_{25} works unless contradicted by evidence
- Issues: What assumptions are reasonable? How to handle contradiction?
-
- Rules with fudge factors:
  - A_{25} /→$_{0.3}$ get there on time
  - *Sprinkler* /→ $_{0.99}$ *WetGrass*
  - *WetGrass* /→ $_{0.7}$ *Rain*
- Issues: Problems with combination, e.g., *Sprinkler* causes *Rain*??
-
- Probability
  - Model agent's degree of belief
  - Given the available evidence,
  - A_{25} will get me there on time with probability 0.04

Probability

Probabilistic assertions summarize effects of

laziness: failure to enumerate exceptions, qualifications, etc.

ignorance: lack of relevant facts, initial conditions, etc.

Subjective (personal, Bayesian) probability (belief):

- Probabilities relate propositions to agent's own state of knowledge

  e.g., $P(A_{25} \mid \text{no reported accidents}) = 0.06$

These are not assertions about the world

Probabilities of propositions change with new evidence:

e.g., $P(A_{25} \mid \text{no reported accidents, 5 a.m.}) = 0.15$

Making decisions under uncertainty

Suppose I believe the following:

$P(A_{25}$ gets me there on time | …) $= 0.04$
$P(A_{90}$ gets me there on time | …) $= 0.70$
$P(A_{120}$ gets me there on time | …) $= 0.95$
$P(A_{1440}$ gets me there on time | …) $= 0.9999$

- Which action to choose?
- Depends on my preferences for missing flight vs. time spent waiting, etc.
  - Utility theory is used to represent and infer preferences
  - Decision theory = probability theory + utility theory

Interpretations of probability

- Sources of uncertainty
  - inherent uncertainty in the physical process;
  - inherent uncertainty at macroscopic level;
  - ignorance;
  - practical omissions;
- Interpretations of probabilities:
  - combinatoric;
  - physical propensities;
  - frequentist;
  - personal/subjectivist;
  - instrumentalist;

$$\lim_{N \to \infty} \frac{N_A}{N} = \lim_{N \to \infty} \hat{p}_N(A) = p(A) ? p(A | \xi)$$

Uncertainty

- .A.Einstein: „God does not play dice..”

  https://arxiv.org/ftp/arxiv/papers/1301/1301.1656.pdf

- Einstein-Podolski-Rosen paradox / Bell Test
- S. Hawking: „Does god play dice?”

  http://www.hawking.org.uk/does-god-play-dice.html

- The BIG Bell Test (Nov30, 2016)
  - http://bist.eu/100000-people-participated-big-bell-test-unique-worldwide-quantum-physics-experiment/

A chronology of uncertain inference

- [1713] Ars Conjectandi (The Art of Conjecture), Jacob Bernoulli
  - **Subjectivist interpretation** of probabilities
- [1718] The Doctrine of Chances, Abraham de Moivre
  - the first textbook on probability theory
  - **Forward predictions**
    - „given a specified number of white and black balls in an urn, what is the probability of drawing a black ball?”
    - his own death
- [1764, posthumous] Essay Towards Solving a Problem in the Doctrine of Chances, Thomas Bayes
  - **Backward questions**: „given that one or more balls has been drawn, what can be said about the number of white and black balls in the urn”
- [1812], Théorie analytique des probabilités, Pierre-Simon Laplace
  - General Bayes rule
- ...
- [1921]: **Correlation and causation**, S. Wright’s diagrams
- [1933]: A. Kolmogorov: *Foundations of the Theory of Probability*

Bayes-omics

- **Thomas Bayes (c. 1702 – 1761)**
- Bayesian probability
- Bayes' rule
- Bayesian statistics
- Bayesian decision
- Bayesian model averaging

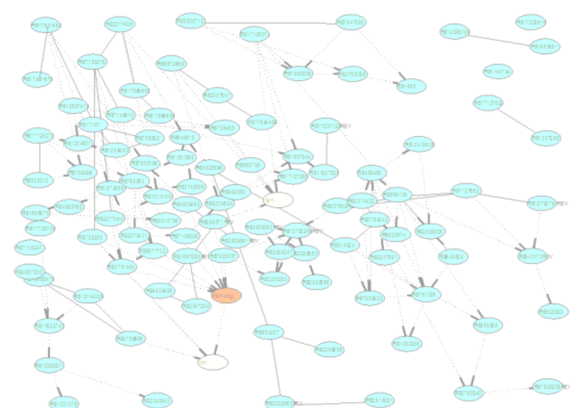
- Bayesian networks
- Bayes factor
- Bayes error
- Bayesian „communication”
- ...

$$p(\text{Model} \mid \text{Data}) \propto p(\text{Data} \mid \text{Model})\, p(\text{Model})$$

$$a^* = \arg\max_i \sum_j U(o_j)\, p(o_j \mid a_i)$$

$$p(prediction \mid data) =$$

$$= \sum_i p(pred. \mid Model_i)\, p(Model_i \mid data)$$



Basic concepts of probability theory

- Joint distribution
- Conditional probability
- Bayes’ rule
- Chain rule
- Marginalization
- General inference
- Independence
  - Conditional independence
  - Independence model

Syntax

- Basic element: random variable
- Similar to propositional logic: possible worlds defined by assignment of values to random variables.

- Boolean random variables
- e.g., *Cavity* (do I have a cavity?)

- Discrete random variables
- e.g., *Weather* is one of <*sunny,rainy,cloudy,snow*>
- Domain values must be exhaustive and mutually exclusive

- Elementary proposition constructed by assignment of a value to a
- random variable: e.g., *Weather* = *sunny*, *Cavity* = *false*
- (abbreviated as $\neg$*cavity*)

- Complex propositions formed from elementary propositions and standard logical connectives e.g., *Weather* = *sunny* $\vee$ *Cavity* = *false*

Syntax

- Atomic event: A complete specification of the state of the world about which the agent is uncertain
-

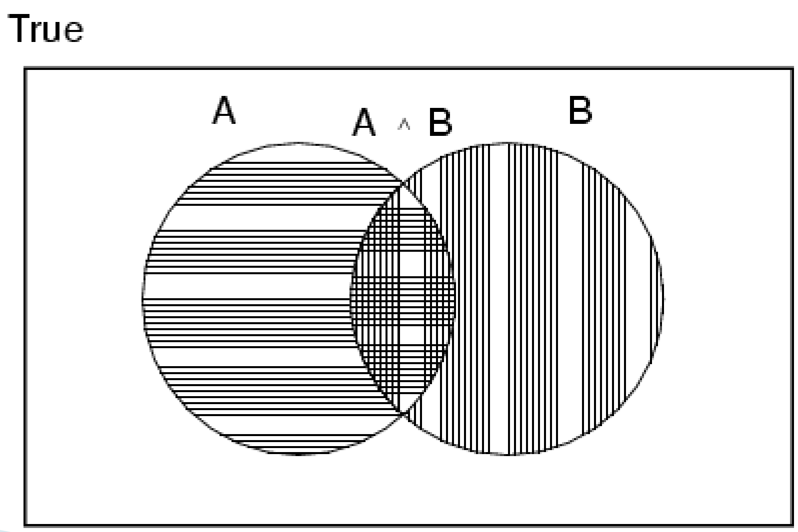
  E.g., if the world consists of only two Boolean variables *Cavity* and *Toothache*, then there are 4 distinct atomic events:

  $Cavity = false \wedge Toothache = false$
  $Cavity = false \wedge Toothache = true$
  $Cavity = true \wedge Toothache = false$
  $Cavity = true \wedge Toothache = true$

- Atomic events are mutually exclusive and exhaustive

Axioms of probability

- For any propositions *A*, *B*
-
    - $0 \le P(A) \le 1$
    - $P(true) = 1$ and $P(false) = 0$
    - $P(A \lor B) = P(A) + P(B) - P(A \land B)$
    -



Joint probability distribution

- Prior or unconditional probabilities of propositions
- e.g., P(*Cavity* = true) = 0.1 and P(*Weather* = sunny) = 0.72 correspond to belief prior to arrival of any (new) evidence
- Probability distribution gives values for all possible assignments:
- **P**(*Weather*) = <0.72,0.1,0.08,0.1> (normalized, i.e., sums to 1)
- Joint probability distribution for a set of random variables gives the probability of every atomic event on those random variables
- **P**(*Weather*,*Cavity*) = a 4 × 2 matrix of values:

Weather =	sunny	rainy	cloudy	snow
Cavity = true	0.144	0.02	0.016	0.02
Cavity = false	0.576	0.08	0.064	0.08

Conditional probability

- Conditional or posterior probabilities
- e.g., $P(cavity \mid toothache) = 0.8$
  i.e., given that *toothache* is all I know
- (Notation for conditional distributions:
- $\mathbf{P}(Cavity \mid Toothache)$ = 2-element vector of 2-element vectors)
- If we know more, e.g., *cavity* is also given, then we have
- $P(cavity / toothache, cavity) = 1$
- New evidence may be irrelevant, allowing simplification, e.g.,
- $P(cavity / toothache, sunny) = P(cavity \mid toothache) = 0.8$
- This kind of inference, sanctioned by domain knowledge, is crucial
-

Conditional probability

- Definition of conditional probability:
- $P(a \mid b) = P(a \wedge b) / P(b)$ if $P(b) > 0$
-
- Product rule gives an alternative formulation:
- $P(a \wedge b) = P(a \mid b) P(b) = P(b \mid a) P(a)$
-
- A general version holds for whole distributions, e.g.,
- $\mathbf{P}(Weather, Cavity) = \mathbf{P}(Weather \mid Cavity)\, \mathbf{P}(Cavity)$
- (View as a set of 4 × 2 equations, not matrix mult.)
-
- Chain rule is derived by successive application of product rule:
- $$\begin{aligned}\mathbf{P}(X_1, \ldots, X_n) &= \mathbf{P}(X_1,\ldots,X_{n-1})\, \mathbf{P}(X_n \mid X_1,\ldots,X_{n-1}) \\ &= \mathbf{P}(X_1,\ldots,X_{n-2})\, \mathbf{P}(X_{n-1} \mid X_1,\ldots,X_{n-2})\, \mathbf{P}(X_n \mid X_1,\ldots,X_{n-1}) \\ &= \ldots \\ &= \pi^n_{i=1}\, \mathbf{P}(X_i \mid X_1, \ldots, X_{i-1})\end{aligned}$$

Bayes rule

An algebraic triviality

$$p(X|Y) = \frac{p(Y|X)p(X)}{p(Y)} = \frac{p(Y|X)p(X)}{\sum_X p(Y|X)p(X)}$$

A scientific research paradigm

$$p(Model|Data) \propto p(Data|Model)p(Model)$$

A practical method for inverting causal knowledge to diagnostic tool.

$$p(Cause|Effect) \propto p(Effect|Cause) \times p(Cause)$$

Inference by enumeration

- Every question about a domain can be answered by the joint distribution.

- Start with the joint probability distribution:
-

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

- For any proposition φ, sum the atomic events where it is true: $P(\phi) = \Sigma_{\omega:\omega \models \phi} P(\omega)$
-

Inference by enumeration

- Start with the joint probability distribution:
-

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	.108	.012	.072	.008
¬ *cavity*	.016	.064	.144	.576

- For any proposition ϕ, sum the atomic events where it is true: $P(\phi) = \Sigma_{\omega:\omega \models \phi} P(\omega)$
-
- $P(\textit{toothache}) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2$
-

Inference by enumeration

- Start with the joint probability distribution:
-

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

- For any proposition ϕ, sum the atomic events where it is true: $P(\phi) = \Sigma_{\omega:\omega \models \phi} P(\omega)$
-
- P(*toothache*) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2
-

Inference by enumeration

- Start with the joint probability distribution:
-

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

- Can also compute conditional probabilities:
-

$$P(\neg cavity \mid toothache) = \frac{P(\neg cavity \wedge toothache)}{P(toothache)} = \frac{0.016+0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4$$

Normalization

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

- Denominator can be viewed as a normalization constant α
-

P(*Cavity* | *toothache*) = α, **P**(*Cavity,toothache*)
= α, [**P**(*Cavity,toothache,catch*) + **P**(*Cavity,toothache*,¬ *catch*)]
= α, [<0.108,0.016> + <0.012,0.064>]
= α, <0.12,0.08> = <0.6,0.4>

General idea: compute distribution on query variable by fixing evidence variables and summing over hidden variables

Inference by enumeration, contd.

Typically, we are interested in the posterior joint distribution of the query variables **Y** given specific values **e** for the evidence variables **E**

Let the hidden variables be **H** = **X** - **Y** - **E**

Then the required summation of joint entries is done by summing out the hidden variables:

$\mathbf{P}(\mathbf{Y} \mid \mathbf{E} = \mathbf{e}) = \alpha \mathbf{P}(\mathbf{Y},\mathbf{E} = \mathbf{e}) = \alpha \Sigma_h \mathbf{P}(\mathbf{Y},\mathbf{E}= \mathbf{e}, \mathbf{H} = \mathbf{h})$

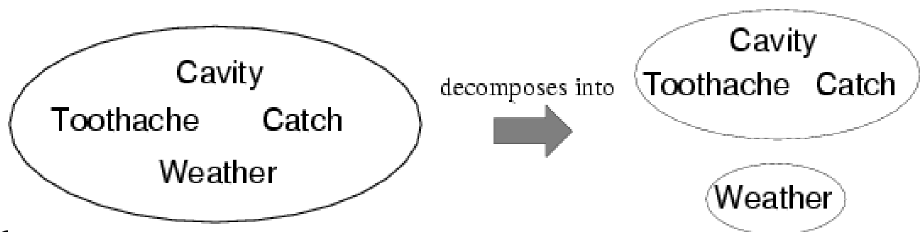
- The terms in the summation are joint entries because **Y, E** and **H** together exhaust the set of random variables
- Obvious problems:
  1. Worst-case time complexity $O(d^n)$ where d is the largest arity
  2. Space complexity $O(d^n)$ to store the joint distribution
  3. How to find the numbers for $O(d^n)$ entries?

Classical vs probabilistic logic: truth and beliefs

P_1	…	P_3	KB	s	pKB	P(query|evidence)
F	F	F	F	T	.01	.1
F	F	T	T	F	.12	.2
F	T	F	F	T	.35	.3
F	T	T	F	F	..	..
T	F	F	F	T	..	..
T	F	T	T	T	…	…
T	T	F	F	T	..	..
T	T	T	F	T	..	..

Independence, conditional independence

- *A* and *B* are independent iff
  P(*A*/*B*) = **P**(*A*) or **P**(*B*/*A*) = **P**(*B*) or **P**(A, B) = **P**(*A*) **P**(*B*)



P(*Toothache, Catch, Cavity, Weather*)
= **P**(*Toothache, Catch, Cavity*) **P**(*Weather*)

- 32 entries reduced to 12; for *n* independent biased coins, $O(2^n) \rightarrow O(n)$
-
- Absolute independence powerful but rare
- *A* and *B* are conditionally independent iff
  P(*A*/*B*) = **P**(*A*) or **P**(*B*/*A*) = **P**(*B*) or **P**(A, B|C) = **P**(A|C) **P**(B|C)

Conditional independence

$I_P(X;Y|Z)$ or $(X \perp\!\!\!\perp Y|Z)_P$ denotes that X is independent of Y given Z defined as follows

for all x,y and z with $P(z)>0$: $P(x;y|z)=P(x|z)\ P(y|z)$

(Almost) alternatively, $I_P(X;Y|Z)$ iff

$P(X|Z,Y)= P(X|Z)$ for all z,y with $P(z,y)>0$.

Other notations: $D_P(X;Y|Z) =\text{def}= \neg\ I_P(X;Y|Z)$

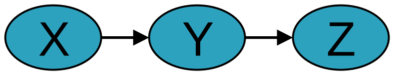
Contextual independence: for not all z.

Direct dependence: $D_P(X;Y|V/\{X,Y\})$

The independence model of a distribution

The independence map (model) M of a distribution P is the set of the valid independence triplets:

$M_P=\{I_{P,1}(X_1;Y_1|Z_1),\ldots, I_{P,K}(X_K;Y_K|Z_K)\}$



If P(X,Y,Z) is a Markov chain, then

$M_P=\{D(X;Y), D(Y;Z), I(X;Z|Y)\}$

Normally/almost always: $D(X;Z)$

Exceptionally: $I(X;Z)$

The semi-graphoid axioms

1. Symmetry: The observational probabilistic conditional independence is symmetric.

$$I_p(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z})\ iff\ I_p(\boldsymbol{Y};\boldsymbol{X}|\boldsymbol{Z})$$

2. Decomposition: Any part of an irrelevant information is irrelevant.

$$I_p(\boldsymbol{X};\boldsymbol{Y}\cup\boldsymbol{W}|\boldsymbol{Z}) \Rightarrow\ I_p(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z})\ and\ I_p(\boldsymbol{X};\boldsymbol{W}|\boldsymbol{Z})$$

3. Weak union: Irrelevant information remains irrelevant after learning (other) irrelevant information.

$$I_p(\boldsymbol{X};\boldsymbol{Y}\cup\boldsymbol{W}|\boldsymbol{Z}) \Rightarrow\ I_p(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z}\cup\boldsymbol{W})$$

4. Contraction: Irrelevant information remains irrelevant after forgetting (other) irrelevant information.

$$I_p(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z})\ and\ I_p(\boldsymbol{X};\boldsymbol{W}|\boldsymbol{Z}\cup\boldsymbol{Y}) \Rightarrow\ I_p(\boldsymbol{X};\boldsymbol{Y}\cup\boldsymbol{W}|\boldsymbol{Z})$$

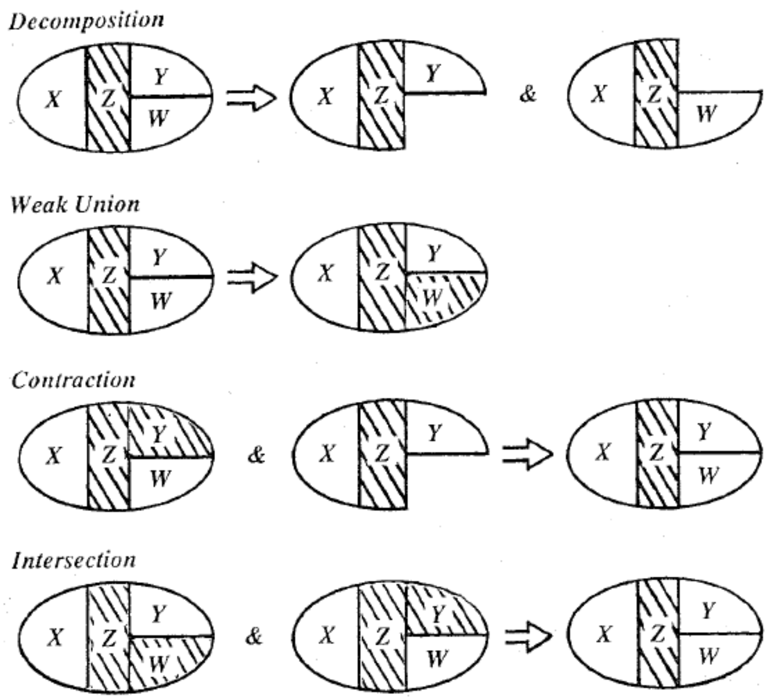
Semi-graphoids (SG): Symmetry, Decomposition, Weak Union, Contraction (holds in all probability distribution). SG is sound, but incomplete inference.

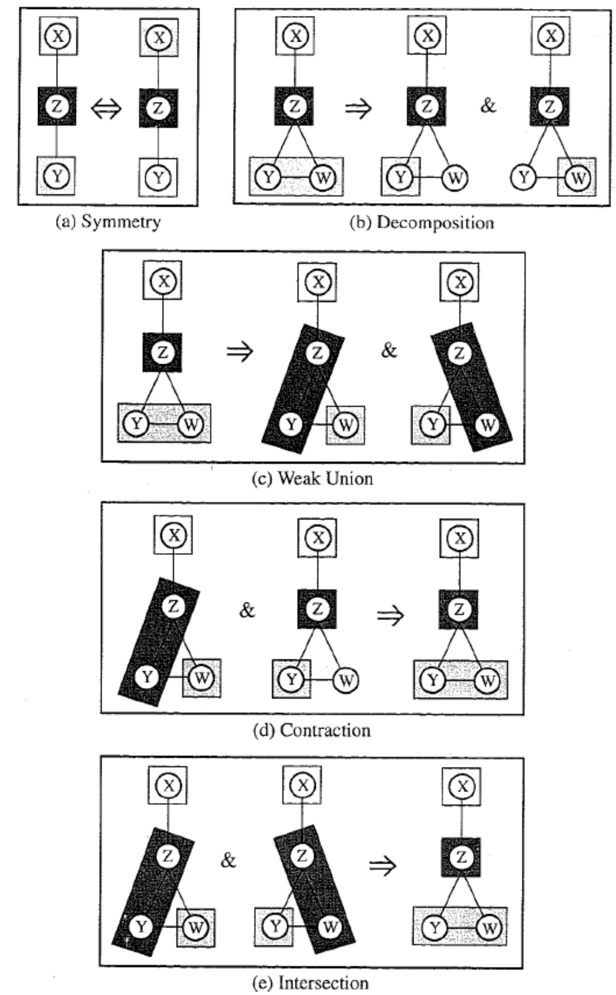
Graphoids

Intersection: Symmetric irrelevance implies joint irrelevance if there are no dependencies.

$$I_p(\mathbf{X};\mathbf{Y}|\mathbf{Z}\cup\mathbf{W})\ and\ I_p(\mathbf{X};\mathbf{W}|\mathbf{Z}\cup\mathbf{Y}) \Rightarrow\ I_p(\mathbf{X};\mathbf{Y}\cup\mathbf{W}|\mathbf{Z})$$

Graphoids: Semi-graphoids+Intersection
(holds only in strictly positive distribution)





J.Pearl: Probabilistic Reasoning in intelligent systems, 1998

Summary

- Probability is a rigorous formalism for uncertain knowledge.
- The subjective/Bayesian interpretation of probabilities avoids the necessity of repeatability.
- Joint probability distribution specifies probability of every atomic event.
- Queries can be answered by summing over atomic events.

- **Suggested reading:**
  - Malakoff: Bayes Offers a `New' Way to Make Sense of Numbers, Science, 1999
  - Efron: Bayes' Theorem in the 21st Century, Science, 2013