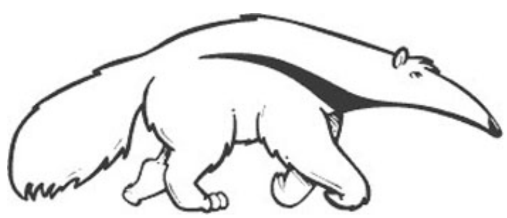
Probability

CS171, Fall 2016

Introduction to Artificial Intelligence

Prof. Alexander Ihler



Reading: R&N Ch 13

BREN:ICS INFORMATION AND COMPUTER SCIENCES

UNIVERSITY of CALIFORNIA IRVINE

Outline

- Representing uncertainty is useful in knowledge bases
  - Probability provides a coherent framework for uncertainty
- Review of basic concepts in probability
  - Emphasis on conditional probability & conditional independence
- Full joint distributions are intractable to work with
  - Conditional independence assumptions allow much simpler models
- Bayesian networks
  - A useful type of structured probability distribution
  - Exploit structure for parsimony, computational efficiency
- **Rational** agents **cannot** violate probability theory

Uncertainty

Let action *At* = leave for airport *t* minutes before flight
Will *At* 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: "*A25* will get me there on time", or
2. leads to conclusions that are too weak for decision making:

"*A25* 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."

"*A1440* should get me there on time but I'd have to stay overnight in the airport."

Uncertainty in the world

- Uncertainty due to
  - Randomness
  - Overwhelming complexity
  - Lack of knowledge
  - …

- Probability gives
  - natural way to describe our assumptions
  - rules for how to combine information

- Subjective probability
  - Relate to agent's own state of knowledge: $P(A25|\text{no accidents}) = 0.05$
  - Not assertions about the world; indicate degrees of belief
  - Change with new evidence: $P(A25 \mid \text{no accidents, 5am}) = 0.20$

Probability

- P(a) is the probability of proposition "a"
  - E.g., P(it will rain in London tomorrow)
  - The proposition a is actually true or false in the real-world
  - P(a) = "prior" or marginal or unconditional probability
  - Assumes no other information is available

- Axioms of probability:
  - 0 <= P(a) <= 1
  - P(NOT(a)) = 1 – P(a)
  - P(true) = 1
  - P(false) = 0
  - P(A OR B) = P(A) + P(B) – P(A AND B)

- Any agent that holds degrees of beliefs that contradict these axioms will act sub-optimally in some cases
  - e.g., de Finetti proved that there will be some combination of bets that forces such an unhappy agent to lose money every time.

- **Rational agents <u>cannot</u> violate probability theory.**

Interpretations of probability

- **Relative Frequency**: *Usually taught in school*
  - P(*a*) represents the frequency that event *a* will happen in repeated trials.
  - Requires event *a* to have happened enough times for data to be collected.

- **Degree of Belief**: *A more general view of probability*
  - P(*a*) represents an agent's degree of belief that event *a* is true.
  - Can predict probabilities of events that occur rarely or have not yet occurred.
  - Does not require new or different rules, just a different interpretation.

- <u>Examples</u>:
  - a = "life exists on another planet"
    - What is P(a)? We will all assign different probabilities
  - a = "California will secede from the US"
    - What is P(a)?
  - a = "over 50% of the students in this class will get A's"
    - What is P(a)?

Concepts of probability

- **Unconditional Probability**
  - **P(a)**, the probability of "a" being true, or **P(a=True)**
  - Does not depend on anything else to be true (**unconditional**)
  - Represents the probability prior to further information that may adjust it (**prior**)
  - Also sometimes "**marginal**" probability (vs. joint probability)

- **Conditional Probability**
  - **P(a|b)**, the probability of "a" being true, given that "b" is true
  - Relies on "b" = true (**conditional**)
  - Represents the prior probability adjusted based upon new information "b" (**posterior**)
  - Can be generalized to more than 2 random variables:
    - e.g. P(a|b, c, d)

- **Joint Probability**
  - **P(a, b) = P(a $\wedge$ b)**, the probability of "a" and "b" both being true
  - Can be generalized to more than 2 random variables:
    - e.g. P(a, b, c, d)

Random variables

- **Random Variable**:
  - Basic element of probability assertions
  - Similar to CSP variable, but values reflect probabilities not constraints.
    - Variable: A
    - Domain: $\{a_1, a_2, a_3\}$ <-- events / outcomes

- Types of Random Variables:
  - Boolean random variables : *{ true, false }*
    - e.g., *Cavity (= do I have a cavity?)*
  - Discrete random variables : one value from a set of values
    - e.g., *Weather is one of {sunny, rainy, cloudy ,snow}*
  - Continuous random variables : a value from within constraints
    - *e.g., Current temperature is bounded by (10°, 200°)*

- Domain values must be exhaustive and mutually exclusive:
  - One of the values must always be the case (**Exhaustive**)
  - Two of the values cannot both be the case (**Mutually Exclusive**)

Random variables

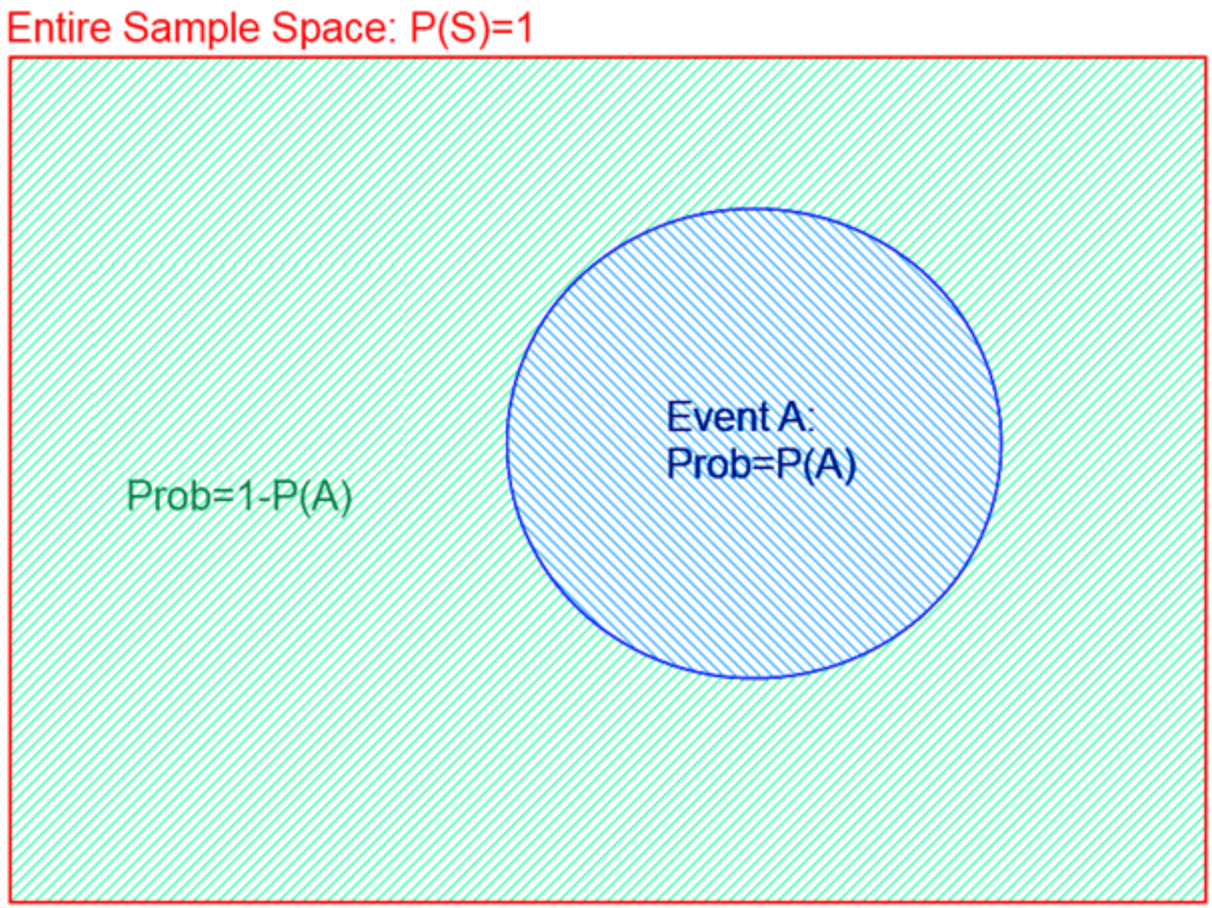
- **Ex**: Coin flip
  - Variable = R, the result of the coin flip
  - Domain = {heads, tails, edge}   <-- must be exhaustive
  - P(R = heads) = 0.4999   }
  - P(R = tails) = 0.4999   } -- must be exclusive
  - P(R = edge) = 0.0002   }

- Shorthand is often used for simplicity:
  - Upper-case letters for variables, lower-case letters for values.
  - e.g.
    - P(a) ≡ P(A = a)
    - P(a|b) ≡ P(A = a | B = b)
    - P(a, b) ≡ P(A = a, B = b)

- Two kinds of probability propositions:
  - **Elementary propositions** are an assignment of a value to a random variable:
    - e.g., *Weather = sunny; Cavity = false (abbreviated as ¬cavity)*
  - **Complex propositions** are formed from elementary propositions and standard logical connectives :
    - e.g., *Cavity = false* ∨ *Weather = sunny*

Probability Space

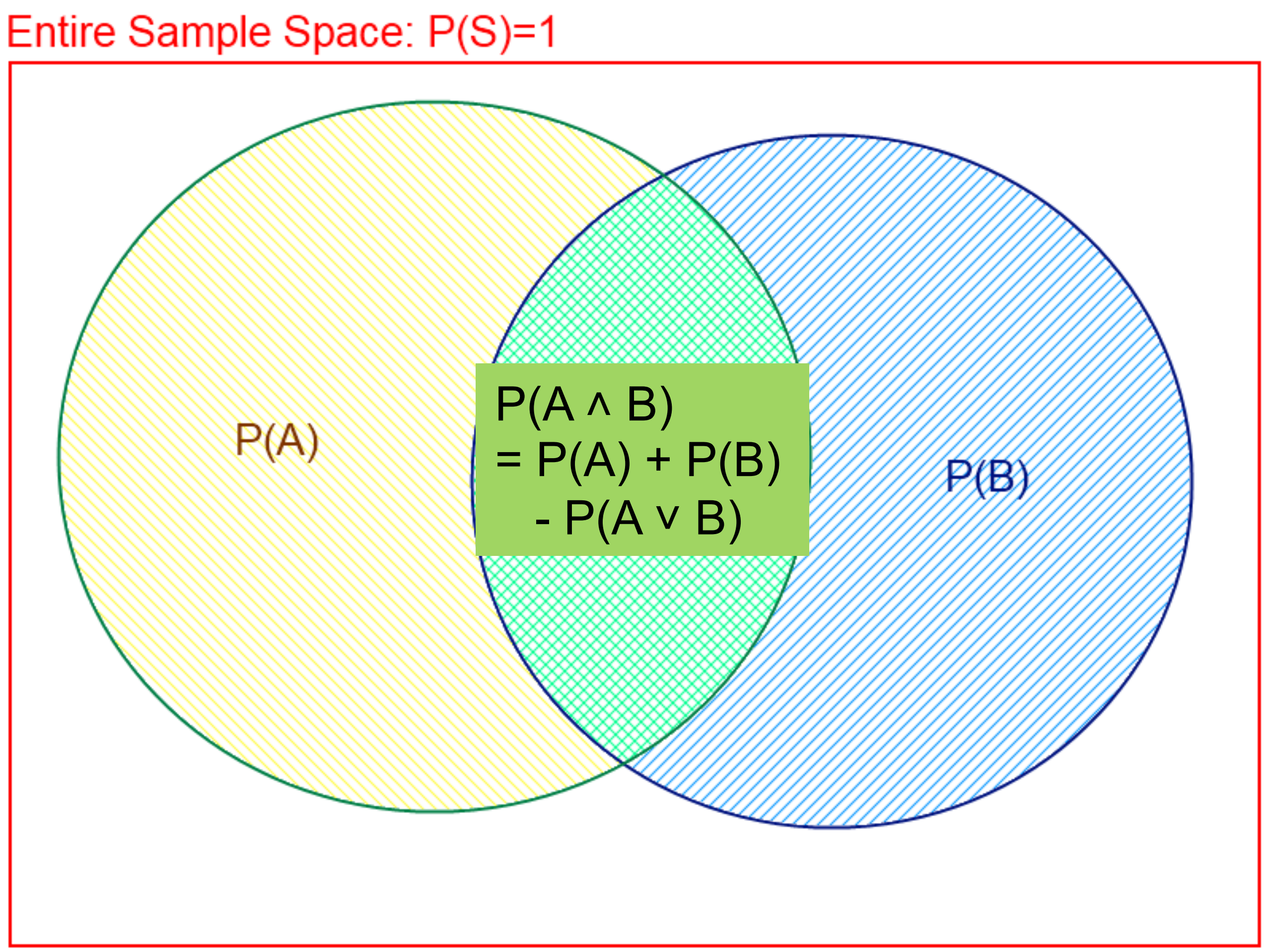
$$P(A) + P(\neg A) = 1$$



Area = Probability of Event

AND Probability

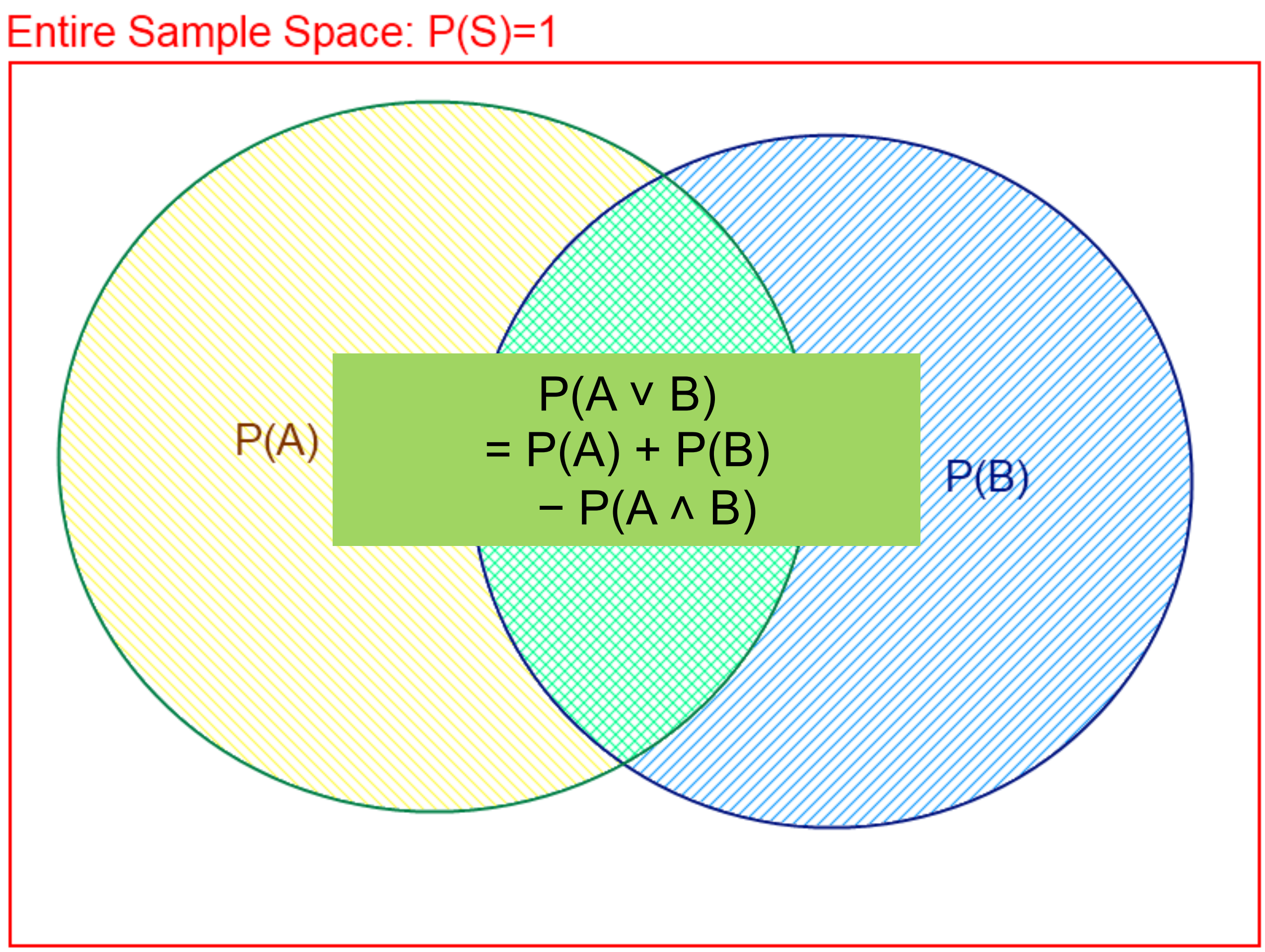
$$P(A, B) = P(A \wedge B) = P(A) + P(B) - P(A \vee B)$$



Area = Probability of Event

OR Probability

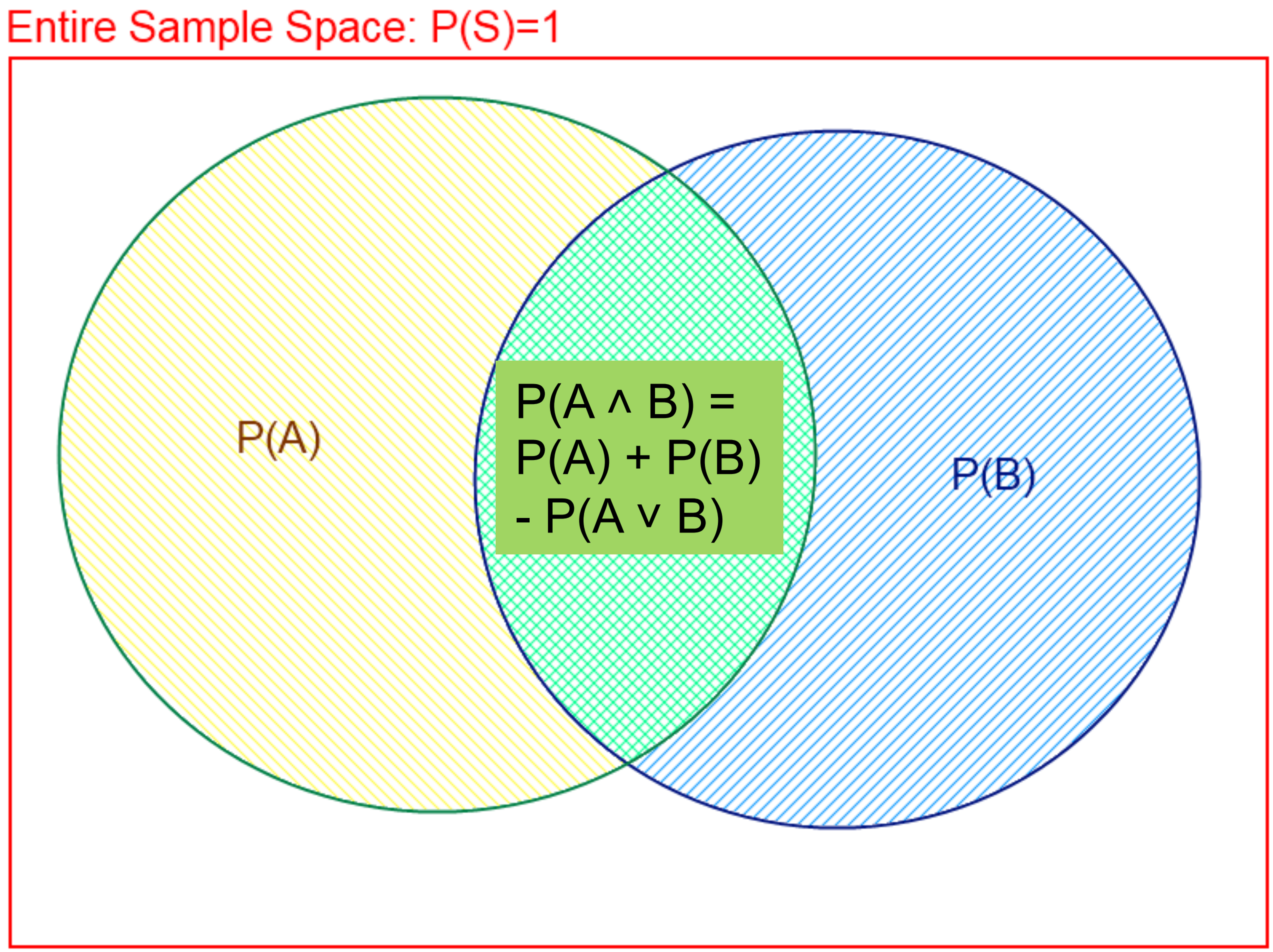
$$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$$



Area = Probability of Event

Conditional Probability

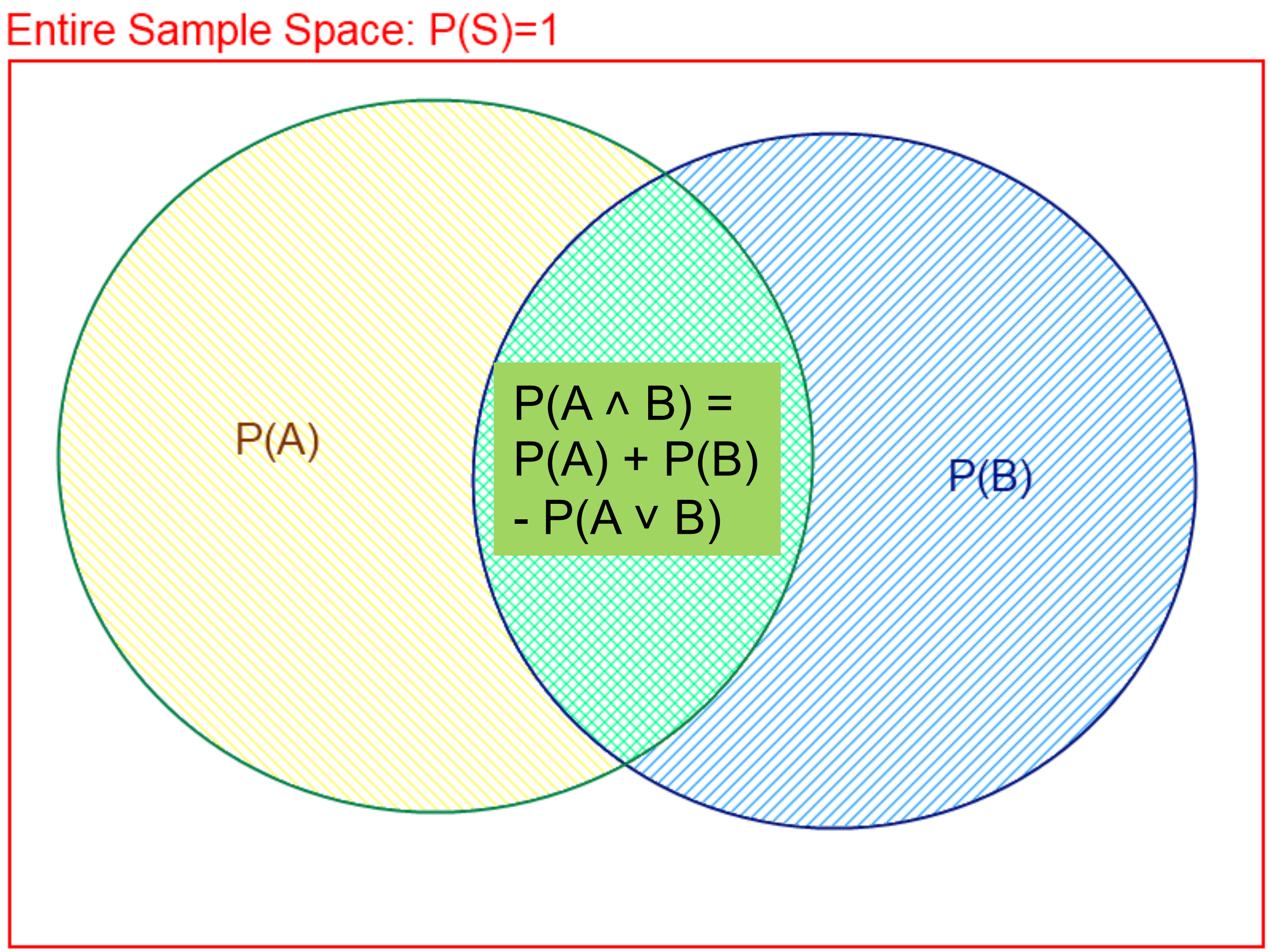
$$P(A \mid B) = P(A, B) / P(B)$$



Area = Probability of Event

Product Rule

$$P(A,B) = P(A|B)\,P(B)$$



Area = Probability of Event

Using the Product Rule

- **Applies to any number of variables:**
  - $P(a, b, c) = P(a, b|c)\, P(c) = P(a|b, c)\, P(b, c)$
  - $P(a, b, c|d, e) = P(a|b, c, d, e)\, P(b, c|d, e)$

- **Factoring**: (AKA **Chain Rule** for probabilities)
  - By the product rule, we can always write:

$$P(a, b, c, \ldots z) = P(a \mid b, c, \ldots. z)\, P(b, c, \ldots z)$$

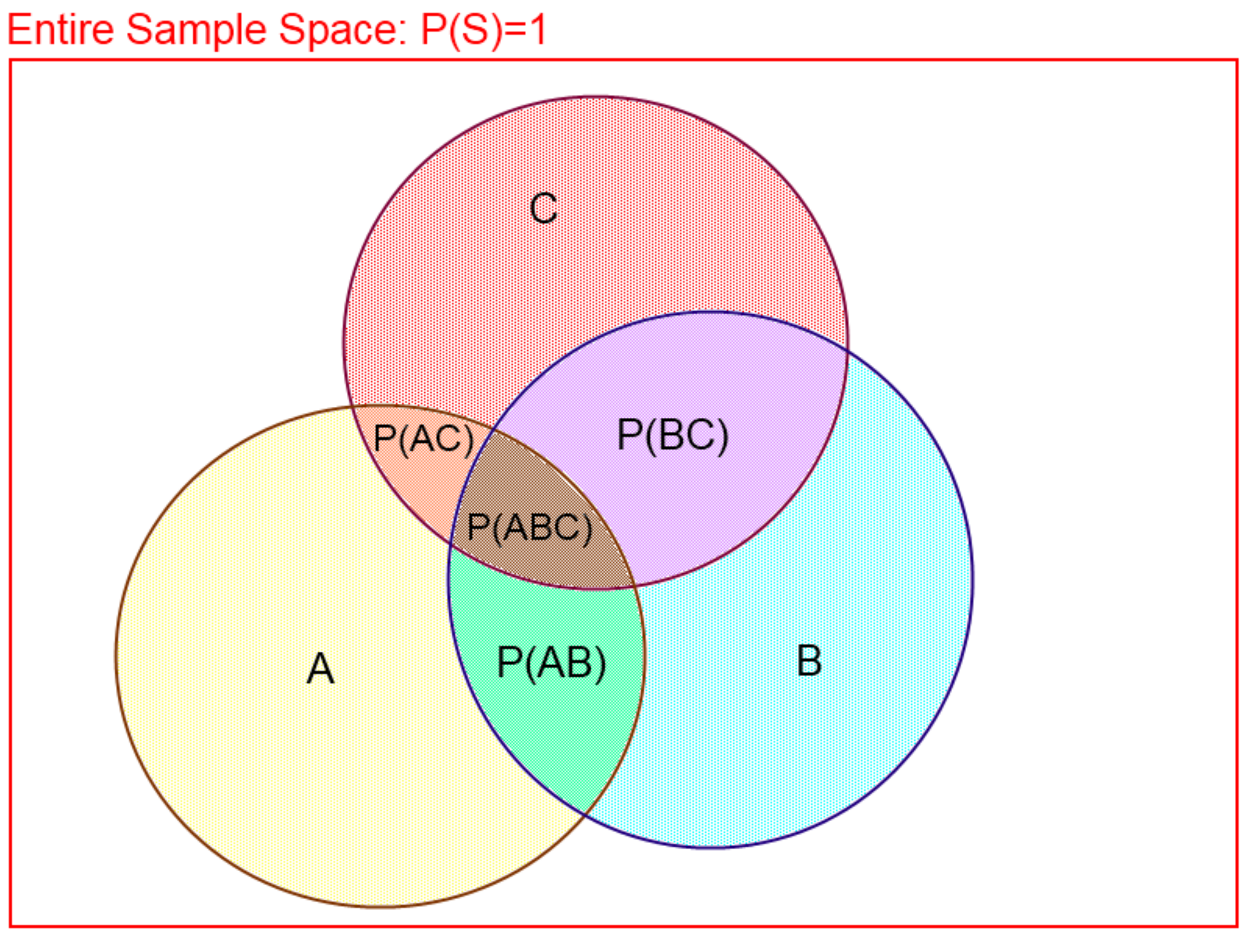
  - Repeatedly applying this idea, we can write:

$$P(a, b, c, \ldots z) = P(a \mid b, c, \ldots. z)\, P(b \mid c, .. z)\, P(c \mid .. z) .. P(z)$$

  - This holds for any ordering of the variables

Sum Rule

$$P(A) = \Sigma_{B,C} P(A,B,C)$$



Area = Probability of Event

Using the Sum Rule

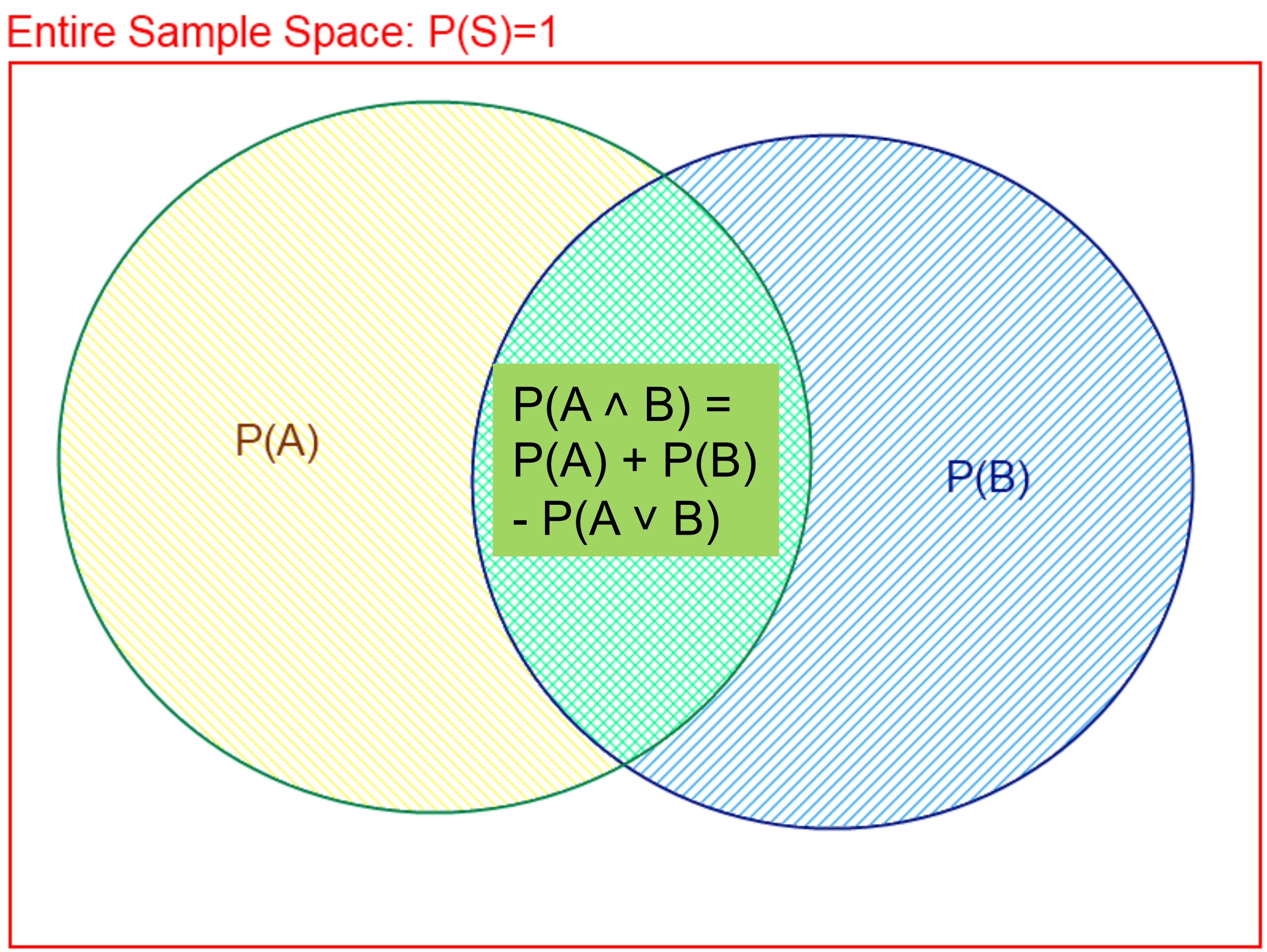
- We can marginalize variables out of any joint distribution by simply summing over that variable:
  - $P(b) = \Sigma_a \Sigma_c \Sigma_d P(a, b, c, d)$
  - $P(a, d) = \Sigma_b \Sigma_c P(a, b, c, d)$

- **For Example**: Determine probability of catching a fish
  - Given a set of probabilities *P(CatchFish, Day, Lake)*
  - Where:
    - *CatchFish =* *{true, false}*
    - *Day =* *{mon, tues, wed, thurs, fri, sat, sun}*
    - *Lake =* *{buel lake, ralph lake, crystal lake}*

  - Need to find P(CatchFish = True):
    - $P(CatchFish = true) = \Sigma_{day} \Sigma_{lake} P(CatchFish = true, day, lake)$

Bayes' Rule

$$P(B|A) = P(A|B)\, P(B) \;/\; P(A)$$



Area = Probability of Event

Derivation of Bayes' Rule

- **Start from Product Rule:**
  - $P(a, b) = P(a|b)\, P(b) = P(b|a)\, P(a)$

- **Isolate Equality on Right Side:**
  - $P(a|b)\, P(b) = P(b|a)\, P(a)$

- **Divide through by P(b):**
  - $P(a|b) = P(b|a)\, P(a) \,/\, P(b)$ <-- **Bayes' Rule**

Summary of probability rules

- **Product Rule**: (aka Chain Rule)
  - **P(a, b) = P(a|b) P(b) = P(b|a) P(a)**
  - Probability of "a" and "b" occurring is the same as probability of "a" occurring given "b" is true, times the probability of "b" occurring.
    - e.g., *P(rain, cloudy) = P(rain | cloudy) * P(cloudy)*

- **Sum Rule**: (aka Law of Total Probability)
  - **P(a) = Σ_b P(a, b) = Σ_b P(a|b) P(b),** where B is any random variable
  - Probability of "a" occurring is the same as the sum of all joint probabilities including the event, provided the joint probabilities represent all possible events.
  - Can be used to "marginalize" out other variables from probabilities, resulting in prior probabilities also being called marginal probabilities.
    - e.g., *P(rain) = $\Sigma_{Windspeed}$ P(rain, Windspeed)*
      where Windspeed = {0-10mph, 10-20mph, 20-30mph, etc.}

- **Bayes' Rule**:
  - **P(b|a) = P(a|b) P(b) / P(a)**
  - Acquired from rearranging the product rule.
  - Allows conversion between conditionals, from P(a|b) to P(b|a).
    - e.g., b = disease, a = symptoms
      More natural to encode knowledge as P(a|b) than as P(b|a).

Example from Russell & Norvig

Joint distributions

- Can fully specify a probability space by constructing a full joint distribution
- Example: dentist
  - T: have a toothache
  - D: dental probe catches
  - C: have a cavity
- Joint distribution
  - Assigns each event (T=t, D=d, C=c) a probability
  - Probabilities sum to 1.0
- Law of total probability:

$$p(C=1) = \sum_{t,p} P(T=t, D=d, C=1)$$

$= 0.008 + 0.072 + 0.012 + 0.108 \quad = 0.20$

  - *Some* value of (T,D) must occur; values disjoint
  - “Marginal probability” of C; “marginalize” or “sum over” T,D

T	D	C	P(T,D,C)
0	0	0	0.576
0	0	1	0.008
0	1	0	0.144
0	1	1	0.072
1	0	0	0.064
1	0	1	0.012
1	1	0	0.016
1	1	1	0.108

Example from Russell & Norvig

The effect of evidence

- Example: dentist
  - T: have a toothache
  - D: dental probe catches
  - C: have a cavity

T	D	C	P(T,D,C)
0	0	0	0.576
0	0	1	0.008
0	1	0	0.144
0	1	1	0.072
1	0	0	0.064
1	0	1	0.012
1	1	0	0.016
1	1	1	0.108

- Recall p(C=1) = 0.20
- Suppose we observe D=0, T=0?

$$p(C=1|D=0,T=0) = \frac{p(C=1,D=0,T=0)}{p(D=0,T=0)}$$

$$= \frac{0.008}{0.576 + 0.008} = 0.012$$

- Observe D=1, T=1?

$$= \frac{0.108}{0.016 + 0.108} = 0.871$$

Called *posterior probabilities*

Example from Russell & Norvig

The effect of evidence

- Example: dentist
  - T: have a toothache
  - D: dental probe catches
  - C: have a cavity

T	D	C	P(T,D,C)
0	0	0	0.576
0	0	1	0.008
0	1	0	0.144
0	1	1	0.072
1	0	0	0.064
1	0	1	0.012
1	1	0	0.016
1	1	1	0.108

- Combining these rules:

$$p(C=1|T=1) = \frac{p(C=1,T=1)}{p(T=1)}$$

$$= \frac{0.012 + 0.108}{0.064 + 0.012 + 0.016 + 0.108} = 0.60$$

$p(T=1) =$ 0.20

Called the *probability of evidence*

Computing posteriors

- Sometimes easiest to normalize last

$$p(C|T=1) \;=\; \frac{1}{p(T=1)}\, p(C,T=1) \;\propto\; p(C,T=1) \;=\; \sum_d p(C,d,T=1)$$

T	D	C	P(T,D,C)
0	0	0	0.576
0	0	1	0.008
0	1	0	0.144
0	1	1	0.072
1	0	0	0.064
1	0	1	0.012
1	1	0	0.016
1	1	1	0.108

Assign T=1

D	C	F(D,C)
0	0	0.064
0	1	0.012
1	0	0.016
1	1	0.108

Sum over D

C	G(C)
0	0.08
1	0.120

Normalize

C	P(C|T=1)
0	0.40
1	0.60

Independence

- X, Y independent:
  - $p(X=x,Y=y) = p(X=x)\ p(Y=y)$ for all x,y
  - Shorthand: $p(X,Y) = P(X)\ P(Y)$
  - Equivalent: $p(X|Y) = p(X)$ or $p(Y|X) = p(Y)$ (if $p(Y), p(X) > 0$)
  - Intuition: knowing X has no information about Y (or vice versa)

Independent probability distributions:

A	P(A)
0	0.4
1	0.6

B	P(B)
0	0.7
1	0.3

C	P(C)
0	0.1
1	0.9

This reduces representation size!

Joint:

A	B	C	P(A,B,C)
0	0	0	.4 * .7 * .1
0	0	1	.4 * .7 * .9
0	1	0	.4 * .3 * .1
0	1	1	…
1	0	0	
1	0	1	
1	1	0	
1	1	1	

Independence

- X, Y independent:
  - p(X=x,Y=y) = p(X=x) p(Y=y)   for all x,y
  - Shorthand: p(X,Y) = P(X) P(Y)
  - Equivalent: p(X|Y) = p(X)   or p(Y|X) = p(Y)   (if p(Y), p(X) > 0)
  - Intuition: knowing X has no information about Y (or vice versa)

Independent probability distributions:

A	P(A)
0	0.4
1	0.6

B	P(B)
0	0.7
1	0.3

C	P(C)
0	0.1
1	0.9

Joint:

A	B	C	P(A,B,C)
0	0	0	0.028
0	0	1	0.252
0	1	0	0.012
0	1	1	0.108
1	0	0	0.042
1	0	1	0.378
1	1	0	0.018
1	1	1	0.162

This reduces representation size!

Note: it is hard to "read" independence
from the joint distribution.
We can "test" for it, however.

Conditional Independence

- X, Y independent given Z
  - $p(X=x,Y=y|Z=z) = p(X=x|Z=z)\, p(Y=y|Z=z)$ for all x,y,z
  - Equivalent: $p(X|Y,Z) = p(X|Z)$ or $p(Y|X,Z) = p(Y|Z)$ (if all > 0)
  - Intuition: X has no additional info about Y beyond Z's

- Example

X = height p(height|reading, age) = p(height|age)

Y = reading ability p(reading|height, age) = p(reading|age)

Z = age

Height and reading ability are dependent (not independent), but are conditionally independent given age

Conditional Independence

- X, Y independent given Z
  - $p(X=x,Y=y|Z=z) = p(X=x|Z=z)\ p(Y=y|Z=z)$ for all x,y,z
  - Equivalent: $p(X|Y,Z) = p(X|Z)$ or $p(Y|X,Z) = p(Y|Z)$
  - Intuition: X has no additional info about Y beyond Z's

- Example: Dentist

Again, hard to "read" from the joint probabilities; only from the conditional probabilities.

Like independence, reduces representation size!

Joint prob:

T	D	C	P(T,D,C)
0	0	0	0.576
0	0	1	0.008
0	1	0	0.144
0	1	1	0.072
1	0	0	0.064
1	0	1	0.012
1	1	0	0.016
1	1	1	0.108

Conditional prob:

T	D	C	P(T|D,C)
0	0	0	0.90
0	0	1	0.40
0	1	0	0.90
0	1	1	0.40
1	0	0	0.10
1	0	1	0.60
1	1	0	0.10
1	1	1	0.60

Conclusions…

- Representing uncertainty is useful in knowledge bases.
- Probability provides a framework for managing uncertainty.
- Using a full joint distribution and probability rules, we can derive any probability relationship in a probability space.
- Number of required probabilities can be reduced through independence and conditional independence relationships
- Probabilities allow us to make better decisions by using decision theory and expected utilities.
- **Rational agents cannot violate probability theory.**