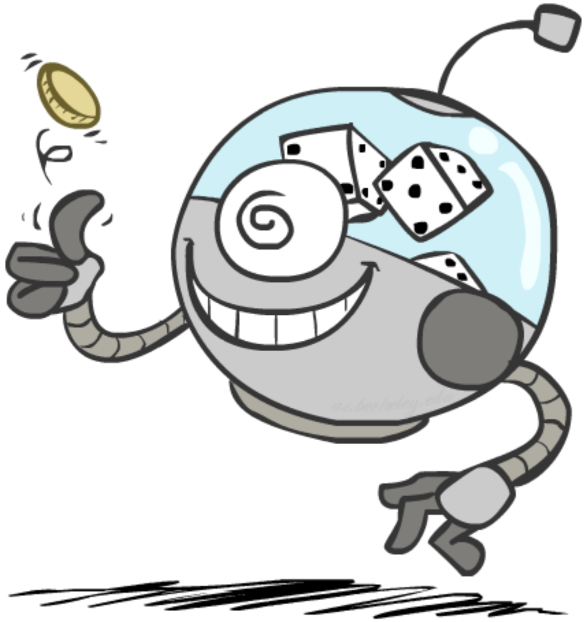
# COMS W4701: Artificial Intelligence

## Lecture 15: Probability Review



Instructor: Tony Dear

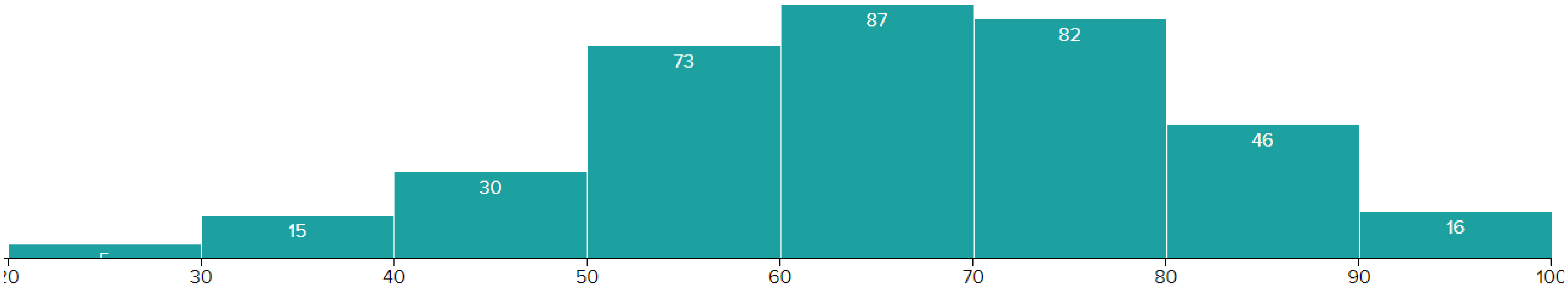
*Lecture materials derived from UC Berkeley's AI course at [ai.berkeley.edu](ai.berkeley.edu)

# Homeworks

- HW4 is now out, due **Friday, Nov 1**
- Similar to HW3, programming-heavy—start early!

- HW3 grades out, regrades by Monday, Oct 28
- Plagiarism

# Midterm




| MEDIAN | MAXIMUM | MEAN | STD DEV |
|---|---|---|---|
| 65.0 | 97.0 | 64.74 | 14.98 |

- Most individual problems distributed similarly to overall distribution
- Outliers: Problem 2 (CSPs) averaged ~50%; 2(c) and 2(d) were each about 33%
- Problem 4(a)-(c) (game trees) were each around 75%; 4(d) (programming) averaged ~50%

# AI Roadmap

- So far: Search, planning, decision-making
- We've considered single-/multi-agent, deterministic/stochastic, episodic/sequential, fully observable, discrete, static situations
- In this class, we will **not** cover any continuous problems

- Recall 90s AI resurgence relied heavily on **probabilistic approaches**
  - Diagnosis, speech and image recognition, tracking, mapping, error correction, etc…
- Why? In the real world, most situations are **partially observable**!
- Agents track their *uncertainty* using *belief states*

# Uncertainty

- **Rationality**—depends on both the return as well as the degree of return
- Translation: How important are goals? How likely are they to be achieved?

- One way to deal with uncertainty: Plan for *all* possible outcomes
- Think expanding a search tree with every possible child

- Better way: *Summarize* uncertainty using probabilities
- Belief states contain both outcomes and likelihoods

# Today

- Discrete random variables, events

- Joint and marginal distributions
- Conditional probabilities

- Product rule, chain rule
- Bayes' theorem and probabilistic inference

# Random Variables

- Recall CSPs: Set of variables $X$, each with a domain $D$
- Similarly, a **random variable** $X: \Omega \rightarrow \mathbb{R}$ is a *function* that maps values in a domain $\Omega$ to a real value (a probability)
- Axioms: $\forall x \;\; P(X = x) \geq 0 \qquad \sum_x P(X = x) = 1$

- Any aspect of the world about which we are uncertain
  - $R$: Is it raining? (Boolean)
  - $N$: How many students predicted to come to class? (Nonnegative integer)
  - $T$: What is the temperature today? (Float, continuous)
  - $L$: Where is a robot on a 2D grid? (Tuples)

# Probability Distributions

- Discrete RVs can be enumerated in a table

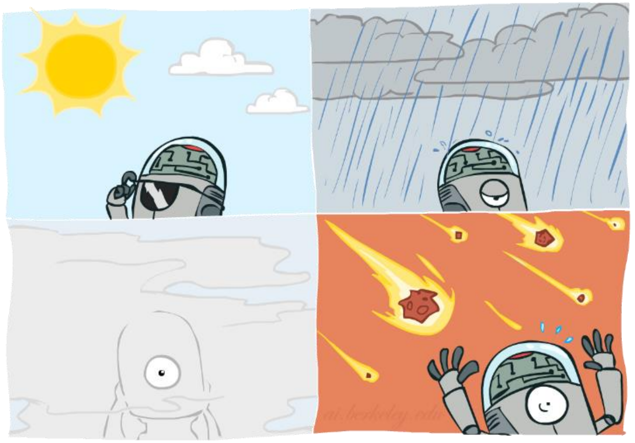
$P(W)$

| W | Pr |
|---|---|
| sun | 0.6 |
| rain | 0.1 |
| fog | 0.3 |
| meteor | 0.0 |

- An **event** $E$ is a *set* of outcomes
- Enumerated by logical propositions

$$P(E) = \sum_{(x_1 \ldots x_n) \in E} P(x_1 \ldots x_n)$$

  - $P(W = sun) = P(sun) = 0.6$
  - $P(W \neq meteor) = P(\sim meteor) = 1.0$
  - $P(rain \text{ OR } fog) = 0.4$



# Joint Probability Distributions

- We can also have probability distributions over *multiple* RVs

$$P(X_1 = x_1, X_2 = x_2, \ldots X_n = x_n)$$

$$P(x_1, x_2, \ldots x_n) \geq 0$$

$$\sum_{(x_1, x_2, \ldots x_n)} P(x_1, x_2, \ldots x_n) = 1$$

- Think of **joint distributions** as *Cartesian product* of RVs
- Size of table = $|X_1| \times |X_2| \times \cdots \times |X_n|$

$$P(T, W)$$

| T | W | Pr |
|---|---|---|
| hot | sun | 0.4 |
| hot | rain | 0.1 |
| cold | sun | 0.2 |
| cold | rain | 0.3 |

- Events over joint distributions:
  - $P(T = hot, W = sun) = P(hot, sun) = 0.4$
  - $P(T = hot, W \neq sun) = P(hot, \sim sun) = 0.1$
  - $P(W = rain) = 0.4$
  - $P(T = hot \text{ OR } W = rain) = P(hot \text{ OR } rain) = 0.8$

# Marginal Probability Distributions

- Given a joint distribution, we can find distributions over *subsets* of RVs
- To do so we can sum out or ***marginalize*** irrelevant RVs
- Check that resultant distribution satisfies probability axioms!

$P(T, W)$

| T | W | Pr |
|---|---|---|
| hot | sun | 0.4 |
| hot | rain | 0.1 |
| cold | sun | 0.2 |
| cold | rain | 0.3 |

$$P(t) = \sum_w P(t, w)$$

$P(T)$

| T | Pr |
|---|---|
| hot | 0.5 |
| cold | 0.5 |

$$P(w) = \sum_t P(t, w)$$

$P(W)$

| W | Pr |
|---|---|
| sun | 0.6 |
| rain | 0.4 |

# Conditional Probabilities

- Marginal probabilities are at least as large as joint probabilities (why?)
- Their ratio is a ***conditional probability***
- Expresses a joint probability within a smaller space

$$P(a|b) = \frac{P(a,b)}{P(b)}$$

$$P(hot|sun) = \frac{P(hot, sun)}{P(sun)} = \frac{0.4}{0.6} = \frac{2}{3}$$

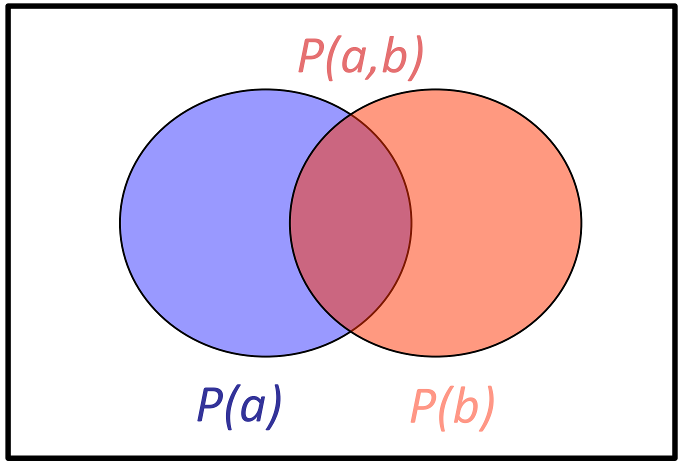
Why is this higher than $P(hot)$?

$$P(sun|hot) = \frac{P(sun, hot)}{P(hot)} = \frac{0.4}{0.5} = \frac{4}{5}$$

Why is this not the same as $P(hot|sun)$?

$P(T, W)$

| T | W | Pr |
|---|---|---|
| hot | sun | 0.4 |
| hot | rain | 0.1 |
| cold | sun | 0.2 |
| cold | rain | 0.3 |



# Conditional Distributions

- Given an observed variable, we can construct a *conditional probability distribution* for all values of an unobserved variable
- When conditioning on a specific event, we end up *normalizing* our joint distribution

$P(T, W)$

| T | W | Pr |
|---|---|---|
| hot | sun | 0.4 |
| hot | rain | 0.1 |
| cold | sun | 0.2 |
| cold | rain | 0.3 |

$P(cold, W)$ →

$P(cold, W)$

| T | W | Pr |
|---|---|---|
| cold | sun | 0.2 |
| cold | rain | 0.3 |

**Normalize** by $P(cold) = 0.5$ →

$P(W|cold)$

| W | Pr |
|---|---|
| sun | 0.4 |
| rain | 0.6 |

$P(T, sun)$ →

$P(T, sun)$

| T | W | Pr |
|---|---|---|
| hot | sun | 0.4 |
| cold | sun | 0.2 |

**Normalize** by $P(sun) = 0.6$ →

$P(T|sun)$

| T | Pr |
|---|---|
| hot | 0.67 |
| cold | 0.33 |

# Conditional Distributions

- While unusual, it is possible to have a "distribution" over *unobserved* variables
- This is not normalization!!!
- Entirely possible to end up with weird tables that do not sum to 1

$P(T, W)$

| T | W | Pr |
|---|---|---|
| hot | sun | 0.4 |
| hot | rain | 0.1 |
| cold | sun | 0.2 |
| cold | rain | 0.3 |

$P(cold, W) \longrightarrow$

$P(cold, W)$

| T | W | Pr |
|---|---|---|
| cold | sun | 0.2 |
| cold | rain | 0.3 |

**Divide** by $P(W)$ $\longrightarrow$

$P(cold|W)$

| W | Pr |
|---|---|
| sun | 0.33 |
| rain | 0.75 |

$P(T, sun) \longrightarrow$

$P(T, sun)$

| T | W | Pr |
|---|---|---|
| hot | sun | 0.4 |
| cold | sun | 0.2 |

**Divide** by $P(T)$ $\longrightarrow$

$P(sun|T)$

| T | Pr |
|---|---|
| hot | 0.8 |
| cold | 0.4 |

# Product Rule

- So far: Joint distributions -> marginal or conditional distributions
- We can also put together a marginal and conditional to recover a joint

$$P(y)P(x|y) = P(x,y)$$

Remember: Marginal RV must be same as the "conditioned" RV

$P(W)$

| W | Pr |
|---|---|
| sun | 0.8 |
| rain | 0.2 |

$P(D|W)$

| D | W | Pr |
|---|---|---|
| wet | sun | 0.1 |
| dry | sun | 0.9 |
| wet | rain | 0.7 |
| dry | rain | 0.3 |

$P(D,W)$

| D | W | Pr |
|---|---|---|
| wet | sun | 0.08 |
| dry | sun | 0.72 |
| wet | rain | 0.14 |
| dry | rain | 0.06 |

# Example

| $Y$ | $Z$ | $Pr(Y,Z)$ |
|---|---|---|
| $+y$ | $+z$ | 0.1 |
| $+y$ | $-z$ | 0.2 |
| $-y$ | $+z$ | 0.3 |
| $-y$ | $-z$ | 0.4 |

| $Y$ | $Z$ | $Pr(+x\|Y,Z)$ |
|---|---|---|
| $+y$ | $+z$ | 0.8 |
| $+y$ | $-z$ | 0.2 |
| $-y$ | $+z$ | 0.5 |
| $+y$ | $-y$ | 0.1 |

Find $\Pr(+x,+y \mid +z)$.

$$\Pr(+x,+y \mid +z) = \frac{\Pr(+x,+y,+z)}{\Pr(+z)}$$

$$= \frac{\Pr(+y,+z)\Pr(+x \mid +y,+z)}{\Pr(+z)} = \frac{0.1 \times 0.8}{0.4} = 0.2$$

$$\Pr(+x,+y \mid +z) = \Pr(+y \mid +z)\Pr(+x \mid +y,+z)$$

$$= \frac{\Pr(+y,+z)}{\Pr(+z)}\Pr(+x \mid +y,+z) = \frac{0.1}{0.4} \times 0.8 = 0.2$$

# Law of Total Probability

- We can combine the product rule with marginalization to find marginal probabilities from conditional probabilities

$$P(x) = \sum_i P(x|y_i)P(y_i) = P(x|y_1)P(y_1) + P(x|y_2)P(y_2) + \cdots + P(x|y_n)P(y_n)$$

$$= \sum_i P(x, y_i) = P(x, y_1) + P(x, y_2) + \cdots + P(x, y_n)$$

$P(W)$

| W | Pr |
|---|---|
| sun | 0.8 |
| rain | 0.2 |

$P(D|W)$

| D | W | Pr |
|---|---|---|
| wet | sun | 0.1 |
| dry | sun | 0.9 |
| wet | rain | 0.7 |
| dry | rain | 0.3 |

$P(D, W)$

| D | W | Pr |
|---|---|---|
| wet | sun | 0.08 |
| dry | sun | 0.72 |
| wet | rain | 0.14 |
| dry | rain | 0.06 |

$P(D)$

| D | Pr |
|---|---|
| wet | 0.22 |
| dry | 0.78 |

# Chain Rule

- The product rule can be extended to more than two RVs!
- Idea: Successively build up larger joint probabilities

$$P(x_1)P(x_2|x_1)P(x_3|x_1,x_2) = P(x_1,x_2)P(x_3|x_1,x_2) = P(x_1,x_2,x_3)$$

- "Proof" of second product: $P(x_3|x_1,x_2) = \dfrac{P(x_1,x_2,x_3)}{P(x_1,x_2)}$

- In general: $$P(x_1,\ldots,x_n) = P(x_1)P(x_2|x_1)\cdots P(x_n|x_1,\ldots,x_{n-1})$$
$$= \prod_i P(x_i|x_1,\ldots,x_{i-1})$$

# Example: Chain Rule

- Show that

$$P(x_1,\ldots,x_i|z_1,\ldots,z_k)P\left(y_1,\ldots,y_j\middle|x_1,\ldots,x_i,z_1,\ldots,z_k\right)=P(x_1,\ldots,x_i,y_1,\ldots,y_j|z_1,\ldots,z_k)$$

$$\frac{\cancel{P(x_1,\ldots,x_i,z_1,\ldots,z_k)}}{P(z_1,\ldots,z_k)}\times\frac{P(x_1,\ldots,x_i,y_1,\ldots,y_j,z_1,\ldots,z_k)}{\cancel{P(x_1,\ldots,x_i,z_1,\ldots,z_k)}}$$

$$=\frac{P(x_1,\ldots,x_i,y_1,\ldots,y_j,z_1,\ldots,z_k)}{P(z_1,\ldots,z_k)}=P(x_1,\ldots,x_i,y_1,\ldots,y_j|z_1,\ldots,z_k)$$

# Probabilistic Inference

- We generally want to *infer* knowledge about some hidden variables given some *evidence*
- $P$(unobserved variables | observed variables)
  - Ex: What is $P(rain|puddle)$?
- Our beliefs generally change with new evidence:
  - $P(rain|puddle, cold) \neq P(rain|puddle)$
- We usually only have $P$(evidence | hidden)
  - Ex: Rain generally leads to puddles (not the other way)

# Bayes' Theorem

- Chain rule takes us from conditional + marginal to a joint probability
- We can also directly convert from one conditional to another

$$P(x,y) = P(x|y)P(y) = P(y|x)P(x) \Rightarrow \boxed{P(x|y) = \frac{P(y|x)}{P(y)}P(x)}$$

- This allows us to "flip" a conditional probability around
- Can be useful for *inferring* or *diagnosing* hidden info given evidence!

$$P(\text{cause}|\text{effect}) = \frac{P(\text{effect}|\text{cause})P(\text{cause})}{P(\text{effect})} \qquad P(\text{hidden} \mid \text{evidence}) = \frac{P(\text{evidence} \mid \text{hidden})P(\text{hidden})}{P(\text{evidence})}$$

# Example: Probabilistic Inference

$$P(\text{cause}|\text{effect}) = \frac{P(\text{effect}|\text{cause})P(\text{cause})}{P(\text{effect})}$$

- Suppose we have two random variables
  - M: meningitis
  - S: stiff neck

$$\left.\begin{aligned} P(+m) &= 0.0001 \\ P(+s|+m) &= 0.8 \\ P(+s|-m) &= 0.01 \end{aligned}\right\}$$

Known probabilities

$$P(+m|+s) = \frac{P(+s|+m)P(+m)}{P(+s)} = \frac{P(+s|+m)P(+m)}{P(+s|+m)P(+m) + P(+s|-m)P(-m)}$$

$$= \frac{0.8 \times 0.0001}{0.8 \times 0.0001 + 0.01 \times 0.999} = 0.008$$

Much smaller than $P(+s|+m)$!