CS 540 Introduction to Artificial Intelligence
Probability

University of Wisconsin-Madison
Fall 2025, Section 3
September 5, 2025

Announcements

- **HW 1:**
  - Short & simple writing assignment
  - Released next Friday (9/12)
  - Due one week later (9/19, 11:59 pm)
- Class roadmap:

Friday **9/5**	**Probability**
Monday 9/8	Linear Algebra
Wednesday 9/10	Statistics
Friday 9/12	Logic
Monday 9/15	NLP

Mostly Foundations

Probability: What is it good for?

Language to express **uncertainty**


Probability in Artificial Intelligence: Prediction

Quantify predictions

$[p(\text{lion}), p(\text{tiger})] = [0.98, 0.02]$

$[p(\text{lion}), p(\text{tiger})] = [0.01, 0.99]$

$[p(\text{lion}), p(\text{tiger})] = [0.43, 0.57]$

Probability in Artificial Intelligence: Generation

Model complex distributions

StyleGAN2 (Karras et al. 2020)
Distribution over space of all possible images.

Probability in ~~Artificial~~ Intelligence: Games

Wisconsin PhD student Ye Yuan finished 5th in 2020 WSOP Main Event

pokernews.com

Jeff Ma, leader of the MIT Blackjack Team and inspiration for the film ***21***

inc.com

Study of probability arose from gambling

Gerolamo Cardano

Liber de ludo aleae (1564)
Book on Games of Chance

Outline

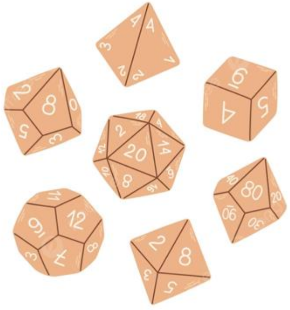
- Basics: definitions, axioms, RVs, joint distributions
- Independence, conditional probability, chain rule
- Bayes’ Rule and Inference

How do we mathematically describe a die?

- We already intuitively understand!

- Introduce formal tools for discussing probability

- These tools “scale” to complex random processes



Basics: Outcomes & Events

- **Outcomes:** possible results of an **experiment**

$$\Omega = \underbrace{\{1, 2, 3, 4, 5, 6\}}_{\text{outcomes}}$$

- **Events:** subsets of outcomes we're interested in

$$\underbrace{\emptyset, \{1\}, \{2\}, \ldots, \{1, 2\}, \ldots, \Omega}_{\text{events}}$$

- Always include $\emptyset, \Omega$

Basics: Probability Distribution

- We have outcomes and events.
- Assign **probabilities**: for each event E, need to have probability $P(E)$
- For a single die:

$$\mathrm{P}(\{1,3,5\}) = \frac{1}{2}$$

equivalent to: $\mathrm{P}(\mathrm{odd}) = \frac{1}{2}$

Basics: **Axioms**

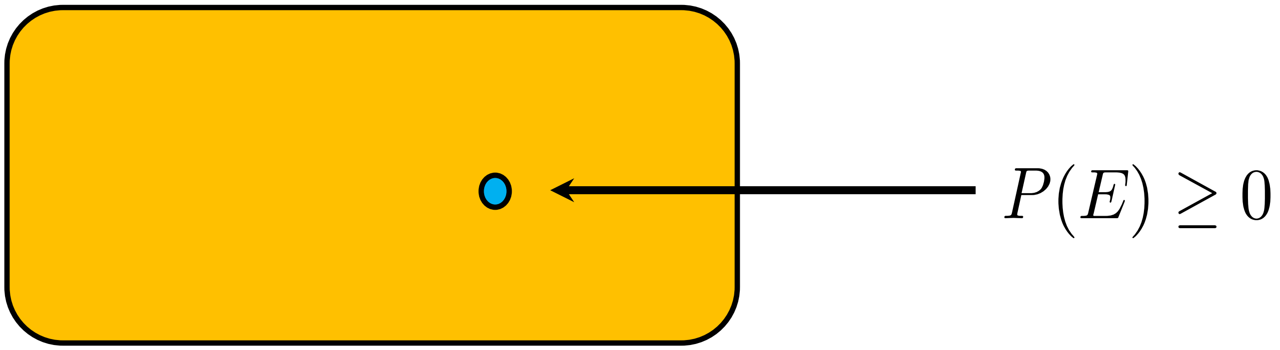
- Rules for probability:
  - For all events E, $P(E) \geq 0$
  - Always, $P(\emptyset) = 0, P(\Omega) = 1$
  - For disjoint events, $P(E_1 \cup E_2) = P(E_1) + P(E_2)$

- Easy to derive other laws. Ex: non-disjoint events

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$$

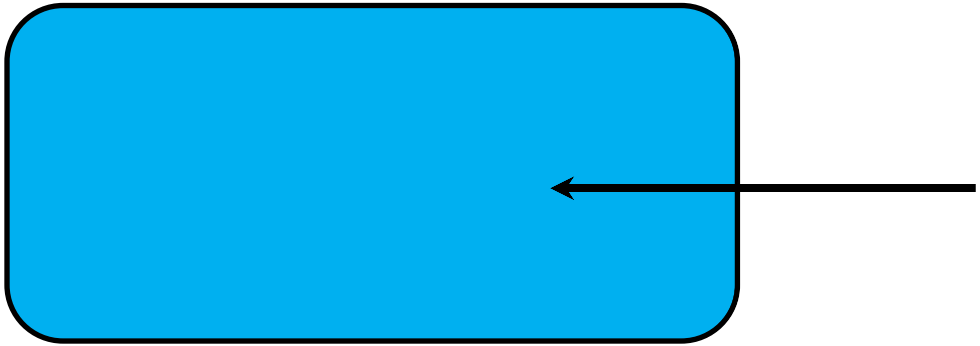
Visualizing the Axioms

- Axiom 1: for all events E, $P(E) \geq 0$



Visualizing the Axioms

- Axiom 2: $P(\emptyset) = 0, P(\Omega) = 1$

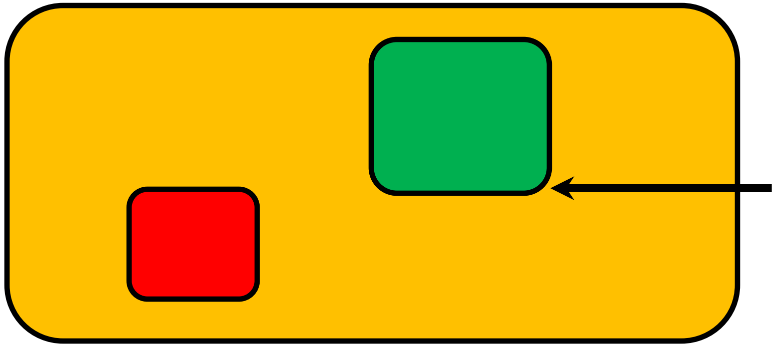


$P(\Omega) = 1$

Visualizing the Axioms: **III**

- Axiom 3: disjoint $$P(E_1 \cup E_2) = P(E_1) + P(E_2)$$

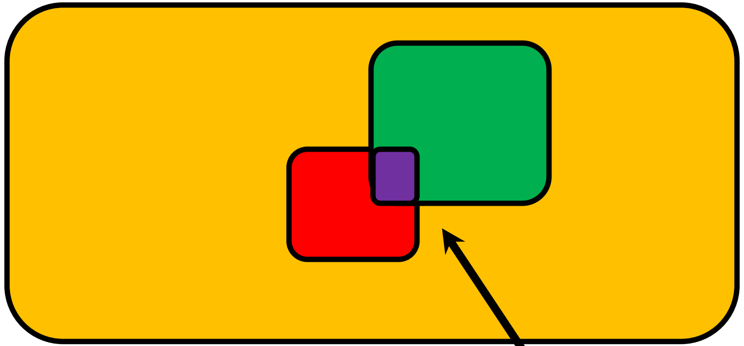
$\boxed{E_1 \textbf{ or } E_2}$



$$P(\color{red}{E_1} \cup \color{green}{E_2}) = P(\color{red}{E_1}) + P(\color{green}{E_2})$$

Visualizing the Axioms

- Also, other laws:



$\boxed{\mathbf{E_1 \text{ and } E_2}}$

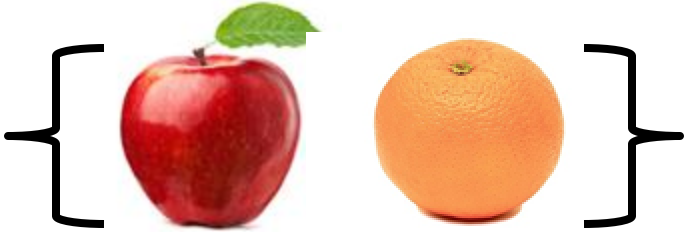
$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$$

Basics: **Random Variables**

- Intuitively: a number X that's random
- Mathematically: map random outcomes to real values

$$X : \Omega \to \mathbb{R}$$

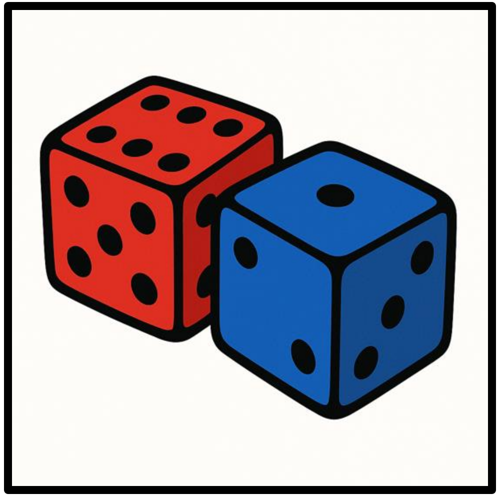
- Why?
  - Previously, everything is a set.
  - Real values are easier to work with



Basics: Random Variables

- Set of outcomes for rolling two dice:

$$\{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),\\ (2,1),(2,2),(2,3),(2,4),(2,5),(2,6),\\ (3,1),(3,2),(3,3),(3,4),(3,5),(3,6),\\ (4,1),(4,2),(4,3),(4,4),(4,5),(4,6),\\ (5,1),(5,2),(5,3),(5,4),(5,5),(5,6),\\ (6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\}$$



- How can we formalize the notion of “the total rolled”?

Basics: Random Variables

- A **random variable** is a function mapping outcomes to numbers

$$f: \Omega \to \mathbb{R}$$

- Confusing name: neither "random" nor a "variable"
- Usually we say: *let random variable X be the total of two dice*

Outcome Ω	Total $f(\Omega)$
(1,1)	2
(1,2)	3
(1,3)	4
(1,4)	5
(1,5)	6
(1,6)	7
(2,1)	3
(2,2)	4
(2,3)	5
(2,4)	6
(2,5)	7
(2,6)	8
(3,1)	4
(3,2)	5
(3,3)	6
(3,4)	7
(3,5)	8
(3,6)	9
(4,1)	5
(4,2)	6
(4,3)	7
(4,4)	8
(4,5)	9
(4,6)	10
(5,1)	6
(5,2)	7
(5,3)	8
(5,4)	9
(5,5)	10
(5,6)	11
(6,1)	7
(6,2)	8
(6,3)	9
(6,4)	10
(6,5)	11
(6,6)	12

Basics: **CDF & PDF**

- Can still work with probabilities:

$$P(X = 3) = p_X(3)$$

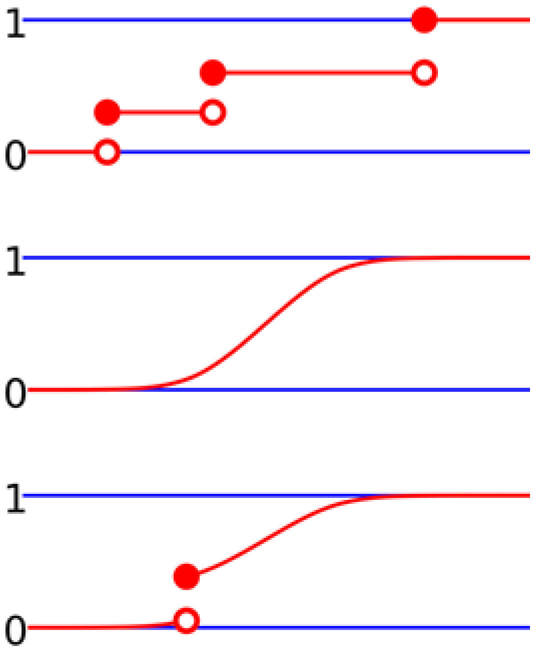
Shorthand for $P(\{(1,2),(2,1)\})$

- Probability density/mass function

$$p_X(x)$$

- Cumulative Distribution Func. (CDF)

$$F_X(x) := P(X \leq x)$$



Wikipedia CDF

Basics: **Expectation & Variance**

- RVs allow us to make useful summaries

- Expectation

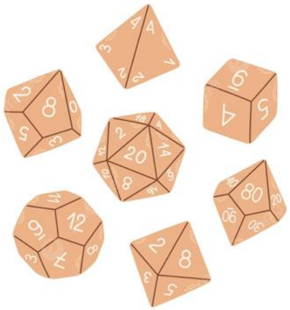
$$\mathbb{E}[X] = \sum_{\text{all values } x} x \cdot P(X = x)$$

- Variance

$$\begin{aligned}\mathrm{Var}[X] &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= \mathbb{E}[(X - \mu)^2]\end{aligned}$$

Basics: **Joint Distributions**

- Move from one variable to several
- Joint distribution: $P(X = a, Y = b)$
  - Why? Work with **multiple** sources of randomness (including possible correlations)

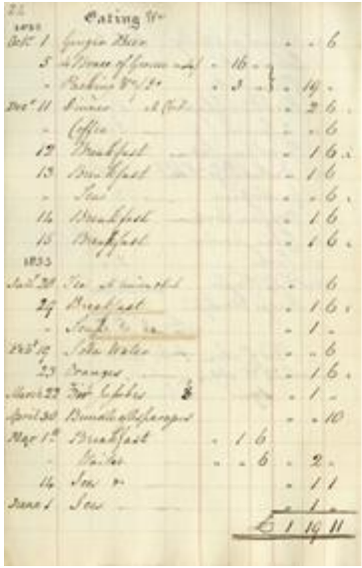


Basics: **Marginal** Probability

- Given a joint distribution $P(X = a, Y = b)$
  - Get the distribution in just one variable:

$$P(X = a) = \sum_b P(X = a, Y = b)$$

  - This is the **marginal** distribution.



Jerry's super blurry camera

- One pixel, 1-bit color sensor (green=trees, white=snow)
- Model T: comes with 1-bit temperature sensor (hot, cold)

- Set of outcomes:

$$\{(\text{green}, \text{hot}), (\text{green}, \text{cold}),$$
$$(\text{white}, \text{hot}), (\text{white}, \text{cold})\}$$

Basics: **Marginal** Probability

$$P(X = a) = \sum_b P(X = a, Y = b)$$

	green	white
hot	150/365	45/365
cold	50/365	120/365

$$[P(\text{hot}), P(\text{cold})] = [\frac{195}{365}, \frac{170}{365}]$$

Probability Tables

- Wrote our distribution as a table

	green	white
hot	150/365	45/365
cold	50/365	120/365

- # of entries? 4
  - If we have n variables with k values, we get k^n entries
  - **Big!** For a 1080p screen, 12 bit color, size of table: $10^{7490589}$
  - No way of writing down all terms

Independence

- Often have unrelated sources of randomness
  - e.g., simultaneously toss a coin and roll a die
- Events $E_1$and E_2 are called **independent** if

$$P(E_1, E_2) = P(E_1) \cdot P(E_2)$$

- Random variables X and Y are called **independent** if, for all a, b

$$P(X = a, Y = b) = P(X = a) \cdot P(Y = b)$$

- Independent: value of two dice rolls
- Not independent: value of two draws from a deck of cards without replacement

Conditional Probability

- The camera sees white. Is it cold outside?

$$P(\,X = a \mid Y = b\,) = \frac{P(X = a, Y = b)}{P(Y = b)}$$

- Our calculation:

	green	white
hot	150/365	45/365
cold	50/365	120/365

$$P(\,cold \mid white\,) = \frac{P(cold, white)}{P(white)} = \frac{120}{45 + 120} = \frac{8}{11} \approx 0.73$$

Conditional Independence

- Events E_1 and E_2 are independent conditioned on E_3 if

$$P(\, E_1, E_2 \mid E_3\,) = P(\, E_1 \mid E_3\,) \cdot P(E_2 \mid E_3)$$

- Similar definition for random variables
- Important throughout course

Chain Rule

- Many events $E_1, E_2, \ldots, E_n$

$$P(E_1, \ldots, E_n) = P(E_1) \times P(E_2 \mid E_1) \times P(E_3 \mid E_2, E_1) \cdots P(E_n \mid E_{n-1}, \ldots, E_1)$$

- Key to language modeling
- Note: may still be big and complicated!
  - If some **conditional independence**, can factor
  - Leads to **probabilistic graphical models**

What have we seen so far?

- Outcomes
- Events
- Probability distribution
- Axioms of probability
- Random variables
- PDF & CDF
- Expectation & variance
- Joint probability
- Marginal probability
- Independence
- Conditional probability
- Conditional independence
- Chain rule

Mathematical tools for discussing and reasoning about randomness

Inference: a core task in artificial intelligence

- Evaluating probabilities:
  - Wake up with a sore throat.
  - Do I have the flu?

- Can we use logic to conclude: $S \rightarrow F$
  - Too strong, doesn't account for uncertainty
- **Inference**: update belief given evidence
  - Calculate $P(F \mid S)$

Bayes' Rule

Theorem: For any events A and B, we have

$$P(\, A \mid B\,) = \frac{P(\, B \mid A\,) \cdot P(A)}{P(B)}$$

Thomas Bayes, c. 1701-1761

Proof: Apply the chain rule two different ways:

$$\begin{aligned} P(A, B) &= P(\, A \mid B\,) \cdot P(B) \\ &= P(\, B \mid A\,) \cdot P(A) \end{aligned}$$

Divide both sides by $P(B)$. $\blacksquare$

Applying Bayes' rule

- Wake up with a sore throat. Do I have the flu?

$$P(\text{ flu } | \text{ sore throat }) = \frac{P(\text{ sore throat } | \text{ flu }) \cdot P(\text{flu})}{P(\text{sore throat})}$$

Bayesian Inference (fancy name for applying Bayes' Rule)

- Terminology:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

- *H* is the hypothesis
- *E* is the evidence

Bayesian Inference (fancy name for applying Bayes' Rule)

- Terminology:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)} \longleftarrow \textbf{Prior}$$

- Prior: estimate of the probability **without** evidence

Bayesian Inference (fancy name for applying Bayes' Rule)

- Terminology:

Likelihood

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

- Likelihood: probability of evidence **given a hypothesis**

Bayesian Inference (fancy name for applying Bayes' Rule)

- Terminology:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

$\uparrow$

Posterior

- Posterior: probability of hypothesis **given evidence**.

Readings

- Vast literature on intro probability and statistics.
- Local classes: **Math/Stat 431**

- **Suggested reading**:

  Probability and Statistics: The Science of Uncertainty,

  Michael J. Evans and Jeff S. Rosenthal

  http://www.utstat.toronto.edu/mikevans/jeffrosenthal/book.pdf

  (Chapters 1-3, excluding "advanced" sections)

Break & Quiz

- **Q 1.1**: We toss a biased coin. If P(heads) = 0.7, then P(tails) = ?
- A. 0.4
- B. 0.3
- C. 0.6
- D. 0.5

Break & Quiz

- **Q 1.1**: We toss a biased coin. If P(heads) = 0.7, then P(tails) = ?
- A. 0.4
- **B. 0.3**
- C. 0.6
- D. 0.5

Break & Quiz

- **Q 1.2**: There are exactly 3 candidates for a presidential election. We know X has a 30% chance of winning, B has a 35% chance. What's the probability that C wins?
- A. 0.35
- B. 0.23
- C. 0.333
- D. 0.8

Break & Quiz

- **Q 1.2**: There are exactly 3 candidates for a presidential election. We know X has a 30% chance of winning, B has a 35% chance. What's the probability that C wins?
- **A. 0.35**
- B. 0.23
- C. 0.333
- D. 0.8

Break & Quiz

- **Q 1.3**: What’s the probability of selecting a black card or a number 6 from a standard deck of 52 cards?
- A. 26/52
- B. 4/52
- C. 30/52
- D. 28/52

Break & Quiz

- **Q 1.3**: What’s the probability of selecting a black card or a number 6 from a standard deck of 52 cards?
- A. 26/52
- B. 4/52
- C. 30/52
- **D. 28/52**

Break & Quiz

Q 2.2: Of a company's employees, 30% are women and 6% are married women. Suppose an employee is selected at random. If the employee selected is a woman, what is the probability that she is married?

A. 0.3
B. 0.06
C. 0.24
D. 0.2