CS 540 Introduction to Artificial Intelligence
Probability

Yingyu Liang
University of Wisconsin-Madison
Sept 14, 2021

Based on slides by Fred Sala

Probability: What is it good for?

- Language to express **uncertainty**

In AI/ML Context

- Quantify predictions

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

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

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

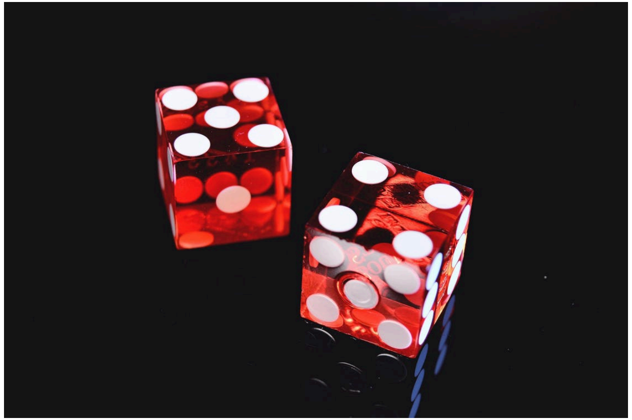
Model Data Generation

- Model complex distributions

StyleGAN2 (Kerras et al ’20)

Outline

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

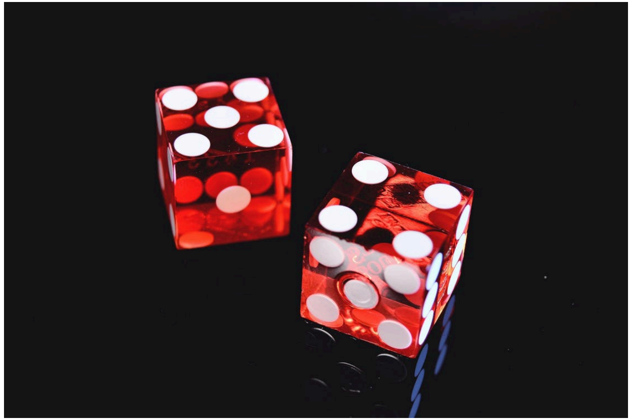


Basics: Outcomes & Events

- Outcomes: possible results of an **experiment**
- **Events:** subsets of outcomes we're interested in

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

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



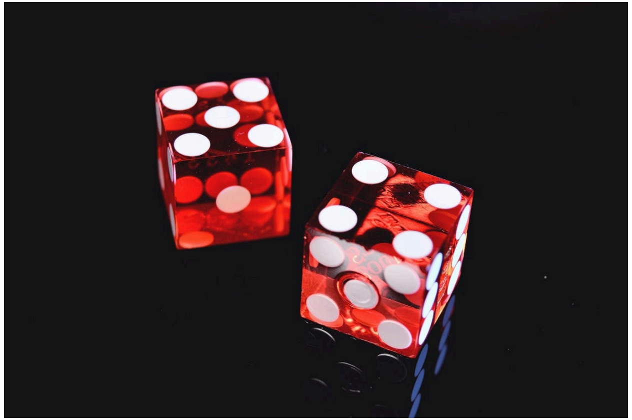
Basics: Outcomes & Events

- Event space can be smaller:

$$\mathcal{F} = \underbrace{\{\emptyset, \{1,3,5\}, \{2,4,6\}, \Omega\}}_{\text{events}}$$

- Two components always in it!

$$\emptyset, \Omega$$



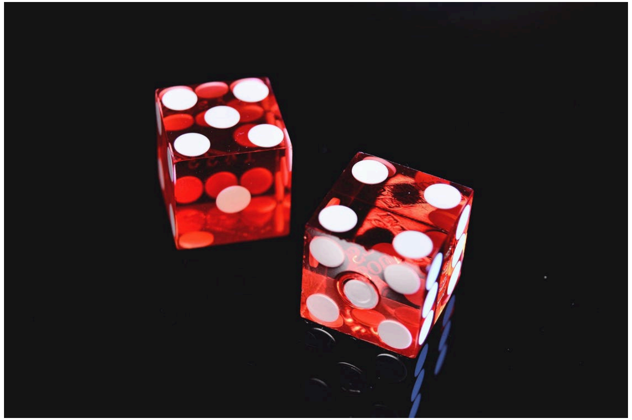
Basics: Probability Distribution

- We have outcomes and events.
- Now assign probabilities $\text{For } E \in \mathcal{F},\ P(E) \in [0,1]$

Back to our example:

$$\mathcal{F} = \underbrace{\{\emptyset, \{1,3,5\}, \{2,4,6\}, \Omega\}}_{\text{events}}$$

$$P(\{1,3,5\}) = 0.2, P(\{2,4,6\}) = 0.8$$



Basics: **Axioms**

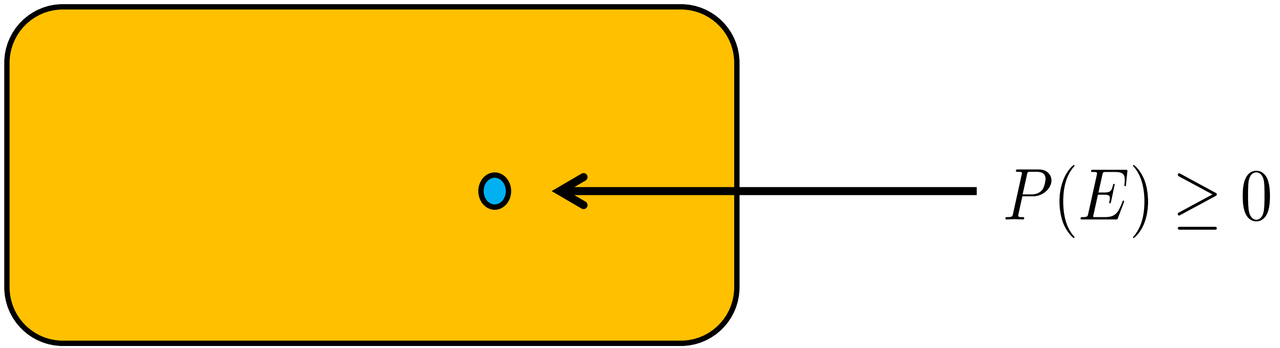
- Rules for probability:
  - For all events $E \in \mathcal{F}, P(E) \geq 0$
  - Always, $P(\emptyset) = 0, P(\Omega) = 1$
  - For disjoint events, $P(E_1 \cup E_2) = P(E_1) + P(E_2)$

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

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

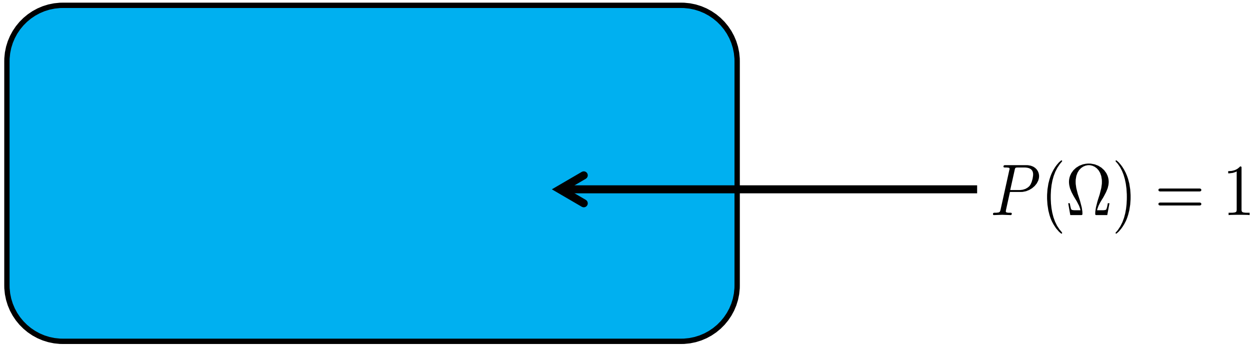
Visualizing the Axioms: I

- Axiom 1: $E \in \mathcal{F}, P(E) \geq 0$



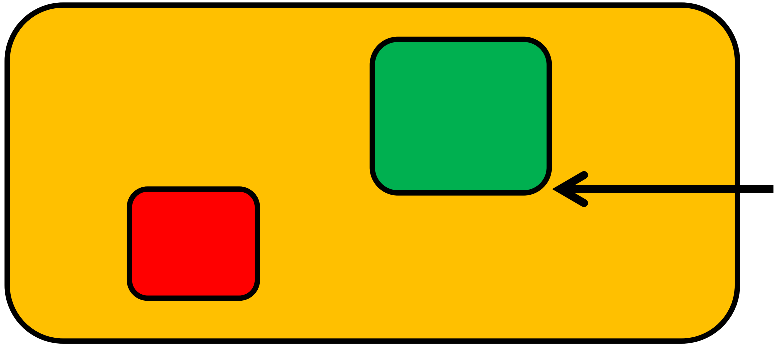
Visualizing the Axioms: **II**

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



Visualizing the Axioms: **III**

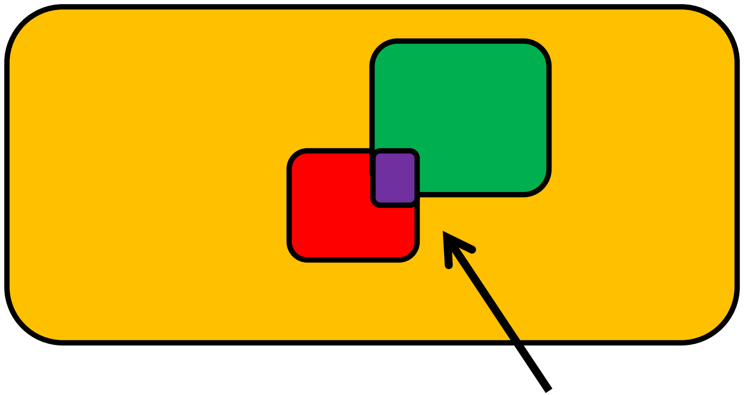
- Axiom 3: disjoint $P(E_1 \cup E_2) = P(E_1) + P(E_2)$



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

Visualizing the Axioms

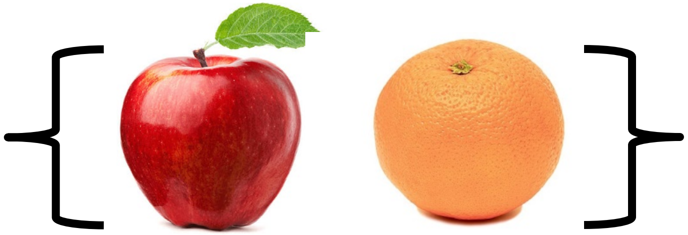
- Also, other laws:



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

Basics: **Random Variables**

- Really, functions
- Map outcomes to real values $X : \Omega \to \mathbb{R}$



- Why?
  - So far, everything is a set.
  - Hard to work with!
  - Real values are easy to work with

Basics: **CDF & PDF**

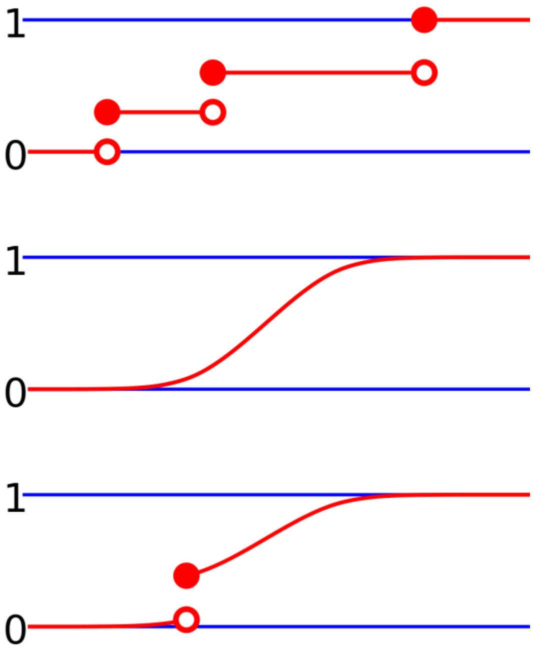
- Can still work with probabilities:

$$P(X=3) := P(\{\omega : X(\omega)=3\})$$

- Cumulative Distribution Func. (CDF)

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

- Density / mass function $p_X(x)$



Wiki CDF

Basics: **Expectation** & **Variance**

- Another advantage of RVs are ``summaries’’
- Expectation: $E[X] = \sum_a a \times P(x = a)$
  - The “average”
- Variance: $Var[X] = E[(X - E[X])^2]$
  - A measure of spread
- Higher moments: other parametrizations

Basics: **Joint Distributions**

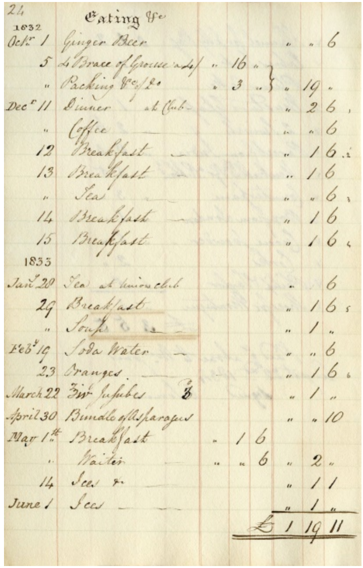
- Move from one variable to several
- Joint distribution: $P(X = a, Y = b)$
  - Why? Work with **multiple** types of uncertainty

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.


24
Eating &c
1832
Octr 1 Ginger Beer 6
Decr 11 Dinner at Club 2 6
Coffee 6
12 Breakfast 1 6
13 Breakfast 1 6
Tea 6
14 Breakfast 1 6
15 Breakfast 1 6
1833
29 Breakfast 1 6
Soup 1
Feb 19 Soda Water 6
23 Oranges 1 6
March 22
April 30 Bundle of Asparagus 10
May 1 Breakfast 1 6
Waiter 6 2
14 Ices 1 1
June 1 Ices 1
£ 1 19 11

Basics: **Marginal** Probability

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

	Sunny	Cloudy	Rainy
hot	150/365	40/365	5/365
cold	50/365	60/365	60/365

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

Probability Tables

- Write our distributions as tables

	Sunny	Cloudy	Rainy
hot	150/365	40/365	5/365
cold	50/365	60/365	60/365

- # of entries? 6.
  - 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

- Independence between RVs:

$$P(X,Y) = P(X)P(Y)$$

- Why useful? Go from k^n entries in a table to $\sim kn$
- Collapses joint into **product** of marginals

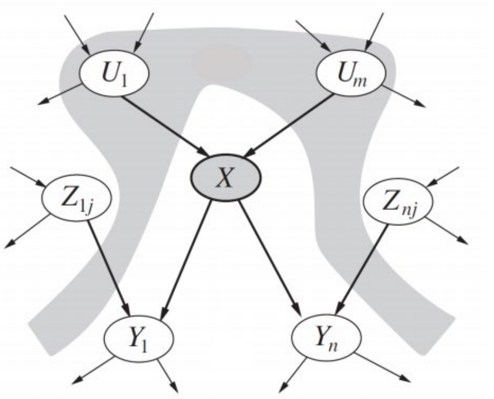
Conditional Probability

- For when we know something,

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

- Leads to **conditional independence**

$$P(X, Y|Z) = P(X|Z)P(Y|Z)$$



Credit: **Devin Soni**

Chain Rule

- Apply repeatedly,

$$P(A_1, A_2, \ldots, A_n)$$
$$= P(A_1)P(A_2|A_1)P(A_3|A_2, A_1) \ldots P(A_n|A_{n-1}, \ldots, A_1)$$

- Note: still big!
  - If some **conditional independence**, can factor!
  - Leads to **probabilistic graphical models**

Reasoning With Conditional Distributions

- Evaluating probabilities:
  - Wake up with a sore throat.
  - Do I have the flu?
- One approach: $S \to F$
  - Too strong.
- **Inference**: compute probability given evidence $P(F|S)$
  - Can be much more complex!

Using **Bayes' Rule**

- Want: $P(F|S)$
- **Bayes' Rule:** $P(F|S) = \frac{P(F,S)}{P(S)} = \frac{P(S|F)P(F)}{P(S)}$
- Parts:
  - $P(S) = 0.1$ Sore throat rate
  - $P(F) = 0.01$ Flu rate
  - $P(S|F) = 0.9$ Sore throat rate among flu sufferers

So: $P(F|S) = 0.09$

Using Bayes' Rule

- Interpretation $P(F|S) = 0.09$
  - Much higher chance of flu than normal rate (0.01).
  - Very different from $P(S|F) = 0.9$
    - 90% of folks with flu have a sore throat
    - But, only 9% of folks with a sore throat have flu

- Idea: **update** probabilities from **evidence**

Bayesian **Inference**

- Fancy name for what we just did. Terminology:

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

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

Bayesian **Inference**

- Terminology:

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

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

Bayesian Inference

- Terminology:

$$P(H|E) = \frac{\underbrace{P(E|H)}_{\text{Likelihood}}P(H)}{P(E)}$$

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

Bayesian Inference

- Terminology:

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

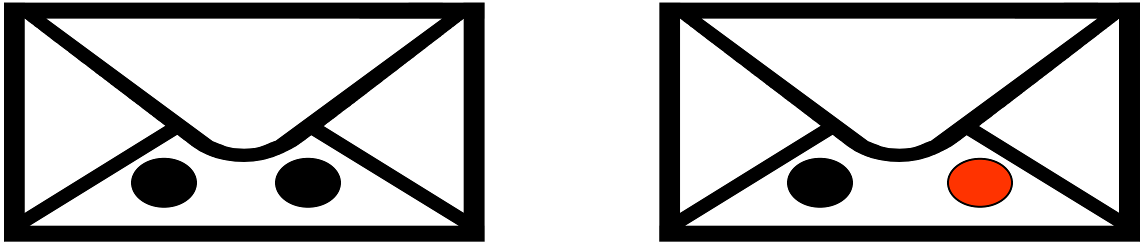
$\uparrow$

Posterior

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

Two Envelopes Problem

- We have two envelopes:
  - E_1 has two black balls, E_2 has one black, one red
  - The **red** one is worth \$100. Others, zero
  - Open an envelope, see one ball. Then, can switch (or not).
  - You see a black ball. **Switch?**



Two Envelopes Solution

- Let's solve it.

$$P(E_1|\text{Black ball}) = \frac{P(\text{Black ball}|E_1)P(E_1)}{P(\text{Black ball})}$$

- Now plug in:

$$P(E_1|\text{Black ball}) = \frac{1 \times \frac{1}{2}}{P(\text{Black ball})}$$

$$P(E_2|\text{Black ball}) = \frac{\frac{1}{2} \times \frac{1}{2}}{P(\text{Black ball})}$$

So switch!