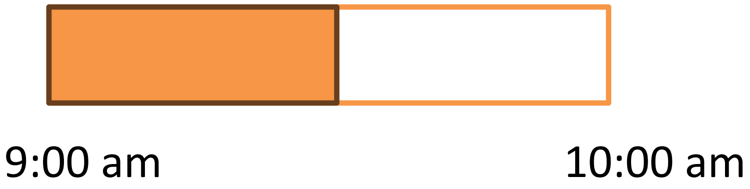
Probability and Information Theory

From the book "Deep Learning“ and Course ECE6001

Random Experiments

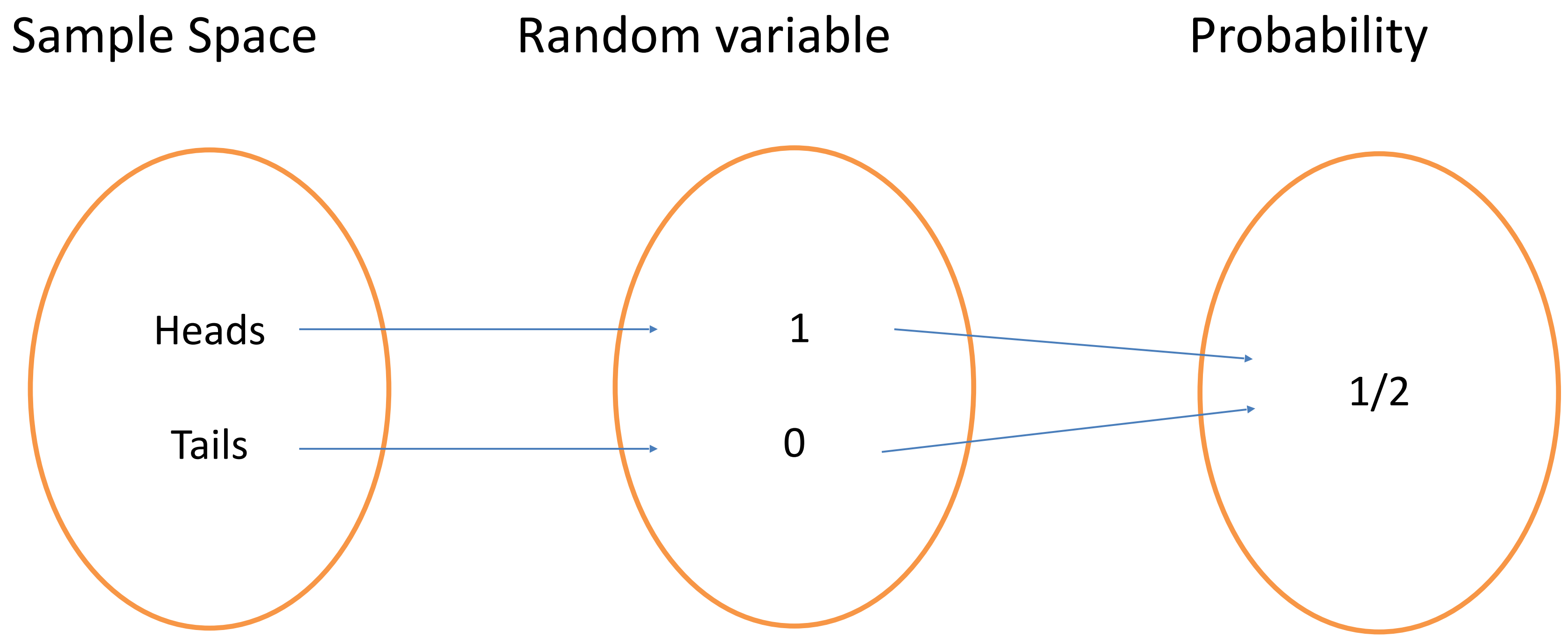
- Single fair coin flip
  - Outcomes: Heads, Tails
  - Discrete
  - Probability of head and probability of tail

- Bus arrival within 9:00 – 10:00 am
  - Continuous: bus can arrive at any time between 9:00 – 10:00 am
  - Probability of bus arrives before 9:30 am with uniform distribution



Random variables

- Random experiment: fair coin flip



A random variable X is a function that assigns real values to outcomes of a random experiment.

X: the number of heads in the outcomes.

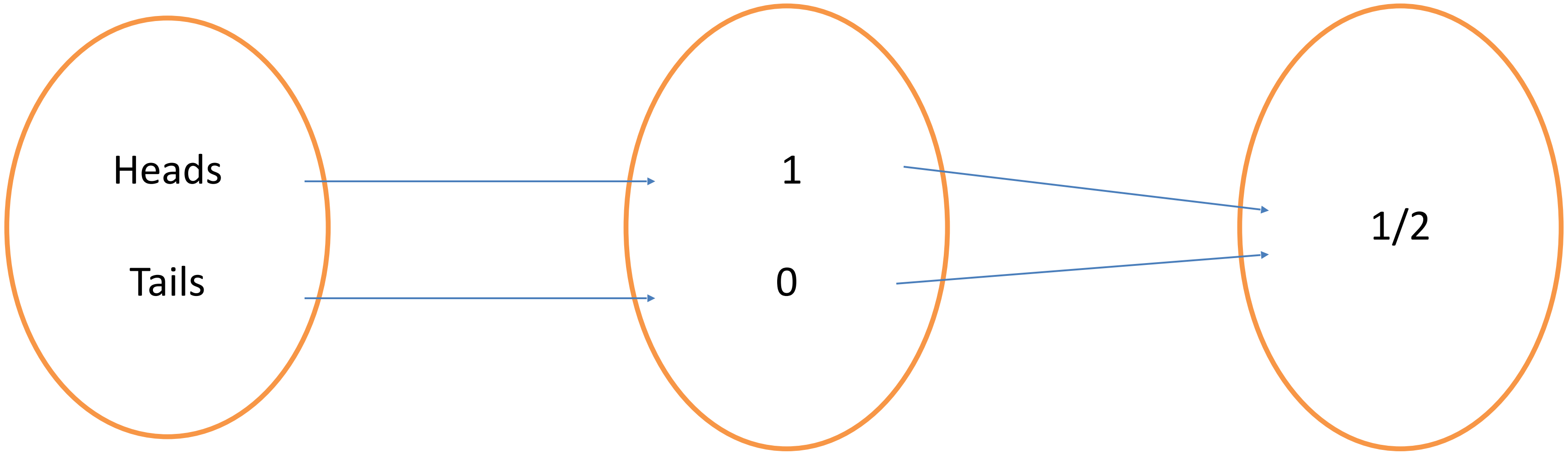
Probability Distributions

- Probability mass functions (PMFs): a probability distribution over discrete variables.
  - Fair coin flips
  - Bernoulli distribution

- Probability density functions (PDFs): a probability distribution over continuous variables.
  - Bus arrival
  - Gaussian Distribution

Bernoulli distribution

- The Bernoulli distribution is a distribution over a single binary random variable.
- Controlled by a single parameter $\phi \in [0,1]$
- $P(X=1)=\phi$
- $P(X=0)=1-\phi$



Maximum Likelihood Estimation (MLE)

- Goal: Estimate unknown parameters using observed data.

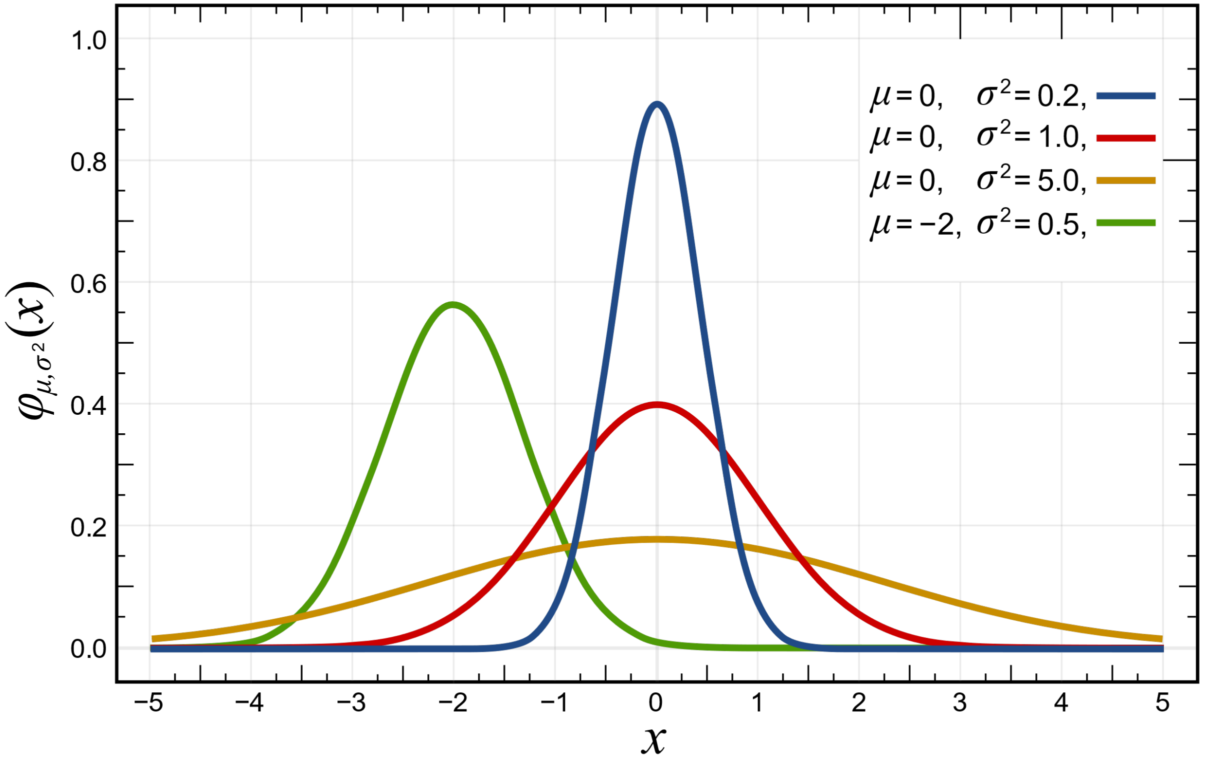
- Example:
- - Toss a coin 10 times, get 7 heads (1) and 3 tails (0).
- - Assume data follows a Bernoulli distribution with unknown p.
- - MLE says: choose p to maximize the likelihood of seeing 7 heads out of 10.

- Answer:
- $\hat{p}$ = number of heads / total tosses = 7 / 10 = 0.7

Gaussian Distribution

- Gaussian Distribution

$$\mathcal{N}(x;\mu,\sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$



Conditional Probability

- The probability of an event occurring given that another event has already occurred.
- We denote the conditional probability that $(Y = y)$ given $(X = x)$ as $P\left(Y = y \middle| X = x\right)$. This conditional probability can be computed with the formula

$$P\left(Y = y \middle| X = x\right) = \frac{P(Y = y, X = x)}{P(X = x)}$$

- Two fair coin flips
  - 1. The probability that both are heads given the first is head
  - 2. The probability that both are heads given at least one is head

Conditional Probability

Outcomes: {HH, HT, TH, TT}

- 1. The probability that both are heads given the first is head
  - $P(\mathrm{HH}) = \frac{1}{4},\ P(\mathrm{HT}, \mathrm{HH}) = \frac{1}{2}, P(\mathrm{HH})/\ P(\mathrm{HT}, \mathrm{HH}) = \frac{1}{2}.$
- 2. The probability that both are heads given at least one is head
  - 2. $P(\mathrm{HH}) = \frac{1}{4},\ P(\mathrm{HT}, \mathrm{TH}, \mathrm{HH}) = \frac{3}{4}, P(\mathrm{HH})/\ P(\mathrm{HT}, \mathrm{HH}) = \frac{1}{3}.$

Independence of Two Events

- Two events are independent is the occurrence or non- occurrence of either one does not affect the probability of the occurrence of the other.

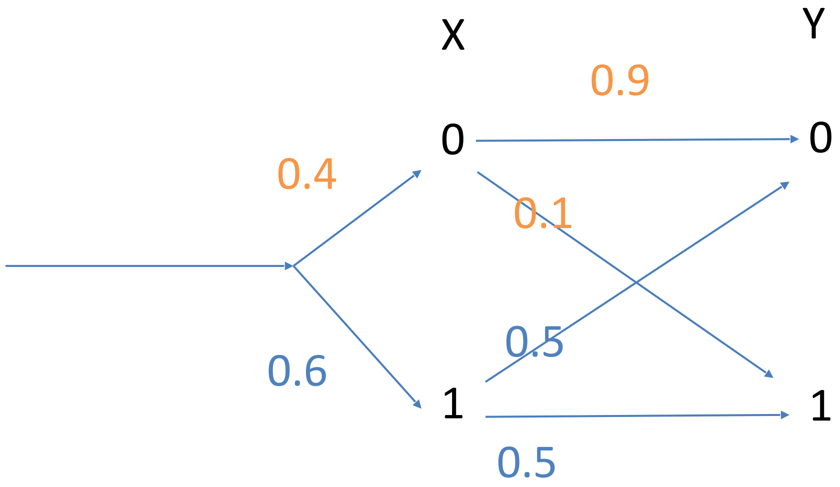
$$P\left(\mathrm{Y} = \mathrm{y}, \mathrm{X} = \mathrm{x}\right) = P(\mathrm{X} = \mathrm{x})\ P\left(\mathrm{Y} = \mathrm{y}\right)$$

- Two fair coin flips: the first is head the second is head.

$$\_\ \mathrm{P(HH)} = \frac{1}{4},\ \mathrm{P(first\ is\ head)} = \frac{1}{2},\ \mathrm{P(second\ is\ head)} = \frac{1}{2}$$

Bayes' Rule

- $P\left(Y = y \middle| X = x\right) = \dfrac{P\left(X = x \middle| Y = y\right)P(Y = y)}{P(X = x)}$
- A random bit is transmitted over the noisy channel



  - $P(X = 0)$ 0.4
  - $P(Y = 0)$ 0.4x0.9+0.6x0.5=0.66
  - $P\left(Y = 0 \middle| X = 0\right)$ 0.9
  - $P\left(X = 0 \middle| Y = 0\right)$ 0.4x0.9/0.66
  - Homework: If $Y = 0$, will predict $X = 0$ or 1? If $Y = 1$, will predict $X = 0$ or 1?

Entropy and Information Theory

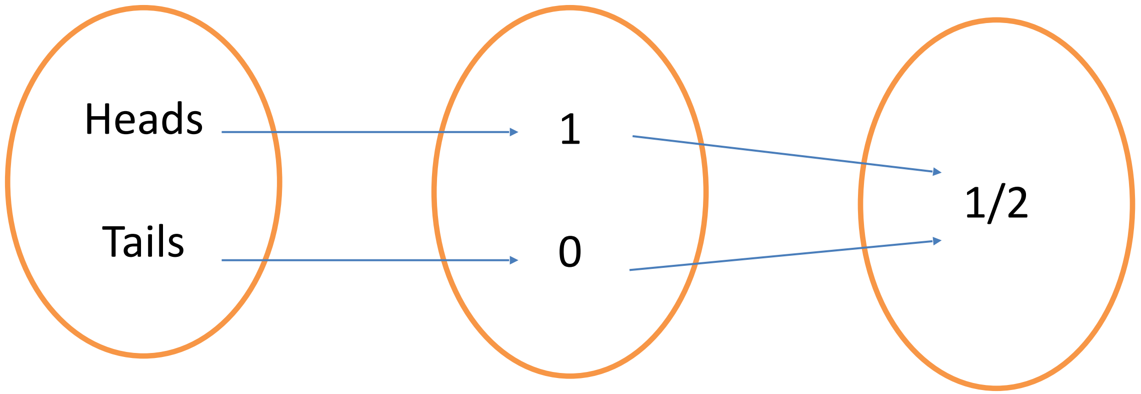
- We define the self-information of an event $\mathrm{X} = \mathrm{x}$ to be

$$I(\mathrm{x}) = -\log P(\mathrm{x})$$

- Shannon entropy use base-2 logarithms

$$H(\mathrm{x}) = \mathbb{E}_{\mathrm{x}\sim P}[I(x)] = -\mathbb{E}_{\mathrm{x}\sim P}[\log P(x)]$$

- One-time fair coin flip



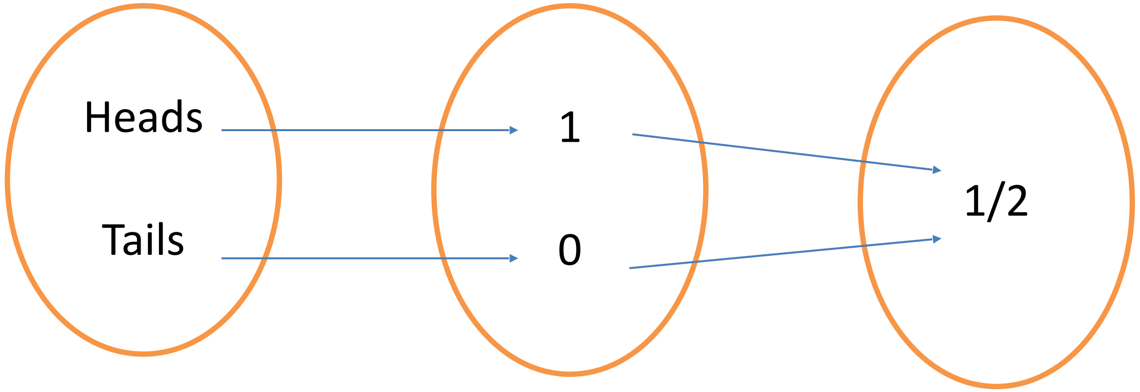
Entropy and Information Theory

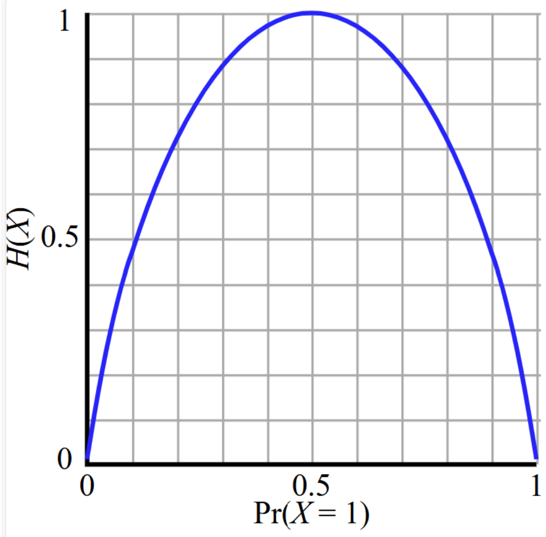
- We define the self-information of an event $\mathrm{X} = \mathrm{x}$ to be

$$I(\mathrm{x}) = -\log P(\mathrm{x})$$

- Shannon entropy use base-2 logarithms

$$H(\mathrm{x}) = \mathbb{E}_{\mathrm{x}\sim P}[I(x)] = -\mathbb{E}_{\mathrm{x}\sim P}[\log P(x)]$$

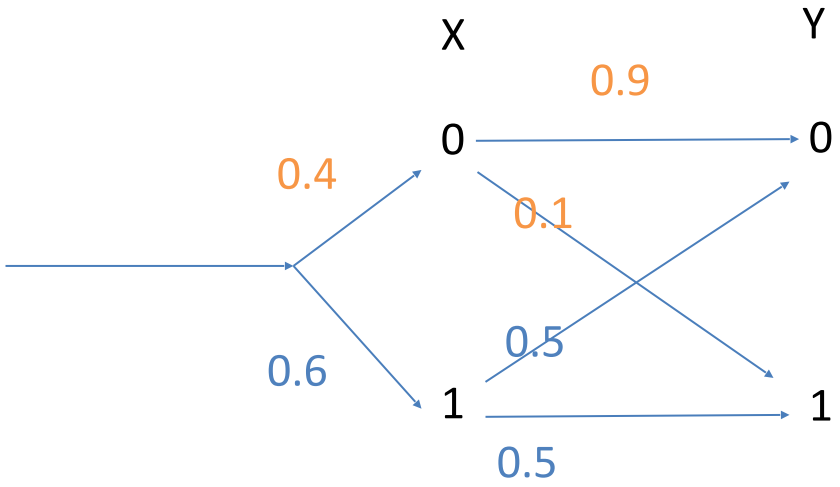




$$H(\mathrm{x}) = -\tfrac{1}{2}\log\tfrac{1}{2} - \tfrac{1}{2}\log\tfrac{1}{2} = 1$$

Bayes' Rule

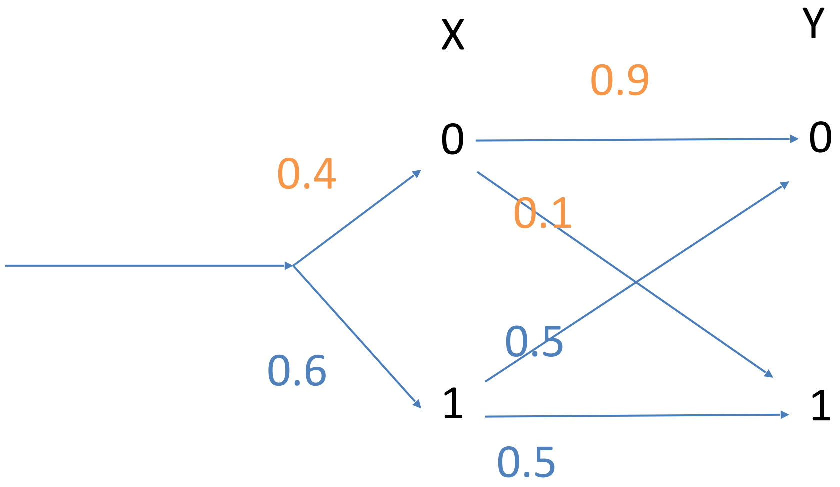
- $P\left(X=0 \middle| Y=0\right) = \dfrac{P\left(Y=0 \middle| X=0\right) P(X=0)}{P(Y=0)}$
- A random bit is transmitted over the noisy channel



  - $P(X=0)$ 0.4
  - $P(Y=0)$ 0.4x0.9+0.6x0.5=0.66
  - $P\left(Y=0 \middle| X=0\right)$ 0.9
  - $P\left(X=0 \middle| Y=0\right)$ 0.4x0.9/0.66 ≈ 0.55 > 0.5
  - Homework: If $Y=0$, will predict $X=0$ or 1? $X=0$

Bayes' Rule

- $$P\left(X=0 \middle| Y=1\right)=\frac{P\left(Y=1 \middle| X=0\right)P(X=0)}{P(Y=1)}$$
- A random bit is transmitted over the noisy channel



  - $P(X=0)$ 0.4
  - $P(Y=1)$ 0.4x0.1+0.6x0.5=0.34
  - $P\left(Y=1 \middle| X=0\right)$ 0.1
  - $P\left(X=0 \middle| Y=1\right)$ 0.1x0.4/0.34 < 0.5
  - Homework: If $Y=1$, will predict $X=0$ or 1? $X=1$

Naive Bayes – Intuition

- Idea: Use probability to classify data.
- Example:
  - - You get an email containing the word 'Win Money'.
  - - From past data, you know these word are common in spam.

- Naive Bayes:
- - For each word, compute how likely it appears in spam vs. non-spam.
- - Multiply the probabilities together (assume independence).
- - Choose the label (spam or not) with the higher overall probability.