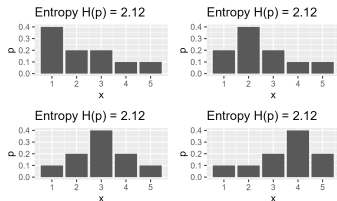
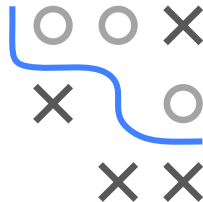


# Introduction to Machine Learning

## Information Theory

### Entropy I

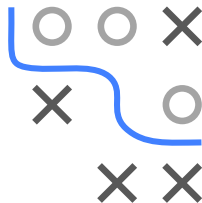
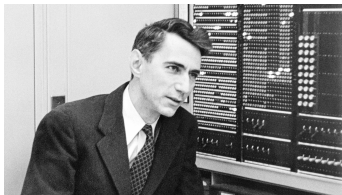


### Learning goals

- Entropy measures expected information for discrete RVs
- Know entropy and its properties

# INFORMATION THEORY

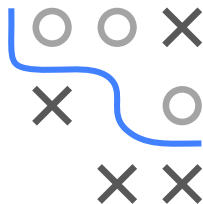
- **Information Theory** is a field of study based on probability theory.
- Foundation was laid by Claude Shannon in 1948; since then been applied in: communication theory, computer science, optimization, cryptography, machine learning and statistical inference.
- Quantify the "amount" of information gained or uncertainty reduced when a random variable is observed.
- Also about storing and transmitting information.



# ENTROPY AS SURPRISAL AND UNCERTAINTY

For a discrete random variable  $X$  with domain  $\mathcal{X} \ni x$  and pmf  $p(x)$ :

$$\begin{aligned} H(X) &:= H(p) = -\mathbb{E}[\log_2(p(X))] &= -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x) \\ &= \mathbb{E} \left[ \log_2 \left( \frac{1}{p(X)} \right) \right] &= \sum_{x \in \mathcal{X}} p(x) \log_2 \frac{1}{p(x)} \end{aligned}$$



Some technicalities first:

- $H$  is actually Greek capital letter **E**ta ( $\eta$ ) for **e**ntropy
- Base of the log simply specifies the unit we measure information in, usually bits (base 2) or 'nats' (base  $e$ )
- If  $p(x) = 0$  for an  $x$ , then  $p(x) \log_2 p(x)$  is taken to be zero, because  $\lim_{p \rightarrow 0} p \log_2 p = 0$ .

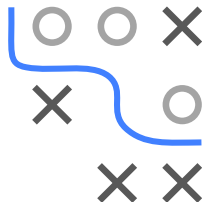


# ENTROPY BASIC PROPERTIES

$$H(X) := H(p) = -\mathbb{E}[\log_2(p(X))] = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x)$$

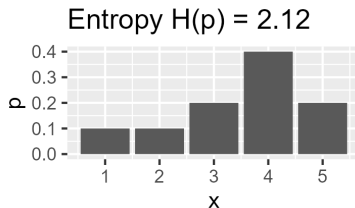
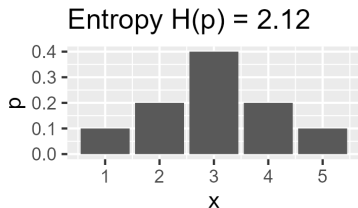
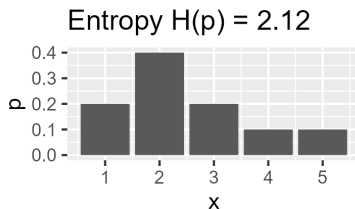
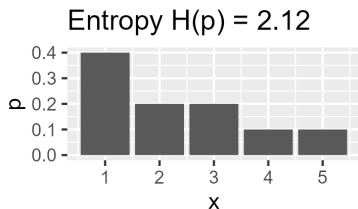
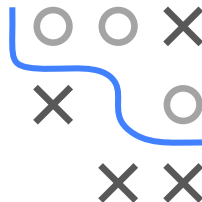
- 1 Entropy is non-negative, so  $H(X) \geq 0$
- 2 If one event has probability  $p(x) = 1$ , then  $H(X) = 0$
- 3 Adding or removing an event with  $p(x) = 0$  doesn't change it
- 4  $H(X)$  is continuous in probabilities  $p(x)$

All these properties follow directly from the definition.



# ENTROPY RE-ORDERING

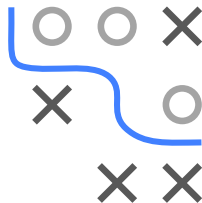
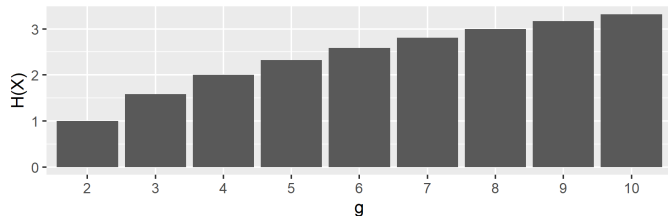
- 5 Symmetry. If the values  $p(x)$  in the pmf are re-ordered, entropy does not change (proof is trivial).



# ENTROPY OF UNIFORM DISTRIBUTIONS

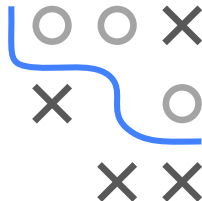
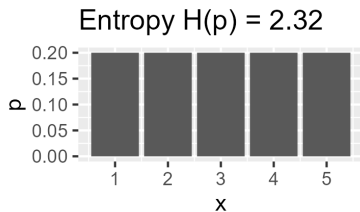
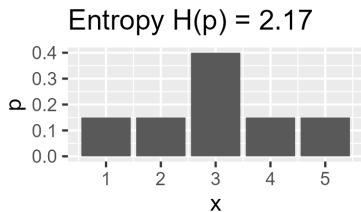
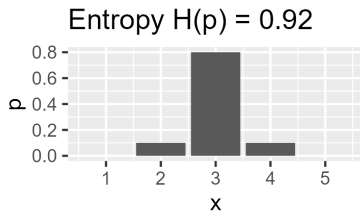
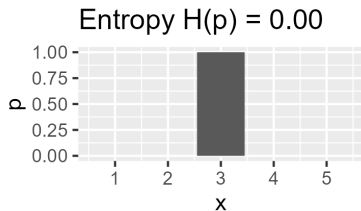
Let  $X$  be a uniform, discrete RV with  $g$  outcomes ( $g$ -sided fair die).

$$H(X) = - \sum_{i=1}^g \frac{1}{g} \log_2 \left( \frac{1}{g} \right) = \log_2 g$$



The more sides a die has, the harder it is to predict the outcome.  
Unpredictability grows *monotonically* with the number of potential outcomes, but at a decreasing rate.

# ENTROPY IS MAXIMAL FOR UNIFORM



- Naive observation:  
Entropy min for 1-point and max for uniform distribution



# ENTROPY IS MAXIMAL FOR UNIFORM

- ⑥ Entropy is maximal for a uniform distribution,  
for domain of size  $g$ :  $H(X) \leq -g \frac{1}{g} \log_2(\frac{1}{g}) = \log_2(g)$ .

**Proof:** So we want to maximize w.r.t. all  $p_i$ :

$$\operatorname{argmax}_{p_1, p_2, \dots, p_g} - \sum_{i=1}^g p_i \log_2 p_i$$

subject to

$$\sum_{i=1}^g p_i = 1$$

