COMP2610/COMP6261 - Information Theory

Lecture 6: Entropy

Mark Reid and **Aditya Menon**

Research School of Computer Science
The Australian National University

THE AUSTRALIAN NATIONAL UNIVERSITY

August 6th, 2014

Last time

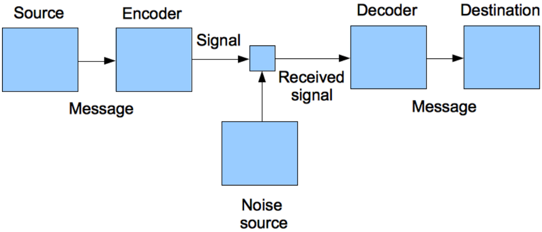
- The Bernoulli and Binomial distributions
- Maximum likelihood estimation
- Bayesian parameter etimation

This time

- Information content and entropy
- Examples and intuition
- Some basic properties of entropy

Outline

Recap: A General Communication System



How informative is a message?

Information Content: Informally

Say that a message comprises a single bit

- Whether or not a coin comes up heads
- Whether or not my favourite horse wins a race

Informally, the amount of information in such a message is:

- How unexpected or "surprising" it is
  - If a coin comes up Heads 99.99% of the time, the message "Tails" is much more informative than "Heads"
  - If I believe my favourite horse will win with 99.99% probability, it is surprising to find out it did not

Information Content: Informally

Say that a message comprises a single bit

- Whether or not a coin comes up heads
- Whether or not my favourite horse wins a race

Informally, the amount of information in such a message is:

- How unexpected or "surprising" it is
  - If a coin comes up Heads 99.99% of the time, the message "Tails" is much more informative than "Heads"
  - If I believe my favourite horse will win with 99.99% probability, it is surprising to find out it did not
- How predictable it is
  - If a coin comes up Heads 99.99% of the time, we can predict the next message as "Heads" and be right most of the time
  - If I believe my favourite horse will win with 99.99% probability, then I believe predicting so to be right most of the time

Information Content: Formally

Intuitively, we measure information of a message in relation to the other messages we could have seen

- For binary messages, we either see 0 or 1
- The message 1 is informative when there is a good chance I might have seen 0

How can we *formalise* and thus *measure* information content?

- Information content of an outcome must depend on its probability
- Information content of a random variable must depend on its probability distribution

Information Content of an Outcome: Definition

Let X be a discrete r.v. with possible outcomes $\mathcal{X}$

- e.g. $\mathcal{X} = \{\texttt{0}, \texttt{1}\}$
- e.g. $\mathcal{X} = \{\texttt{Yes}, \texttt{No}, \texttt{Maybe}\}$

Let $p(x)$ denote the probability of outcome $x \in \mathcal{X}$

The information content of an outcome $x \in \mathcal{X}$ is:

$$h(x) = \log_2 \frac{1}{p(x)}$$

Information Content of an Outcome: Properties

The information content of x grows as $p(x)$ shrinks

- Outcomes that are rare are deemed to contain more information

Choice of logarithm basis is arbitrary

- If we use $\log_2$ we measure information in *bits*

What about other functions of $p(x)$, e.g. $\frac{1}{p(x)^2} - 1$?

Entropy of a Random Variable: Definition

Let X be a discrete r.v. with possible outcomes $\mathcal{X}$.

The entropy of the random variable X is the average information content of the outcomes:

$$
\begin{aligned}
H(X) &= \mathbb{E}_x\left[h(x)\right] \\
&= \sum_x p(x) \cdot \log_2 \frac{1}{p(x)} \\
&= -\sum_x p(x) \log_2 p(x)
\end{aligned}
$$

where we define $0 \log 0 \equiv 0$, as $\lim_{p \to 0} p \log p = 0$.

- Non-negativity:

$$0 \le p(x) \le 1 \to \log \frac{1}{p(x)} \ge 0$$

$$\sum_x p(x) \log \frac{1}{p(x)} \ge 0$$

$$H(X) \ge 0$$

Entropy of a Random Variable

Some Basic Properties

- Non-negativity:

$$0 \leq p(x) \leq 1 \rightarrow \log \frac{1}{p(x)} \geq 0$$

$$\sum_x p(x) \log \frac{1}{p(x)} \geq 0$$

$$H(X) \geq 0$$

- Change of base:

$$H_b(X) = -\sum_x p(x) \log_b p(x)$$

$$= \sum_x p(x) \log_a p(x) \log_b a$$

$$H_b(X) = \log_b a \; H_a(X)$$

  - If we use $\log_2$ the units are called *bits*
  - If we use natural logarithm the units are called *nats*

Unrolling the Definition

The entropy of X is

$$H(X) = -\sum_{x} p(x) \log_2 p(x).$$

Pick a random outcome x, and see how large its probability is

- Average information content of each outcome

Does not depend on the values of the outcomes

- Only on their probabilities

What Does Entropy "Mean"?

Entropy does match some intuitive properties of our informal notion of "information content"

- Rare outcomes provide more information

But other functions also seem plausible, e.g. $G(X) = \sum_x \frac{1}{p(x)}$

We will see some examples where our definition of entropy arises naturally

Entropy of a Random Variable
Example 1 — Bernoulli Distribution

Let $X \in \{0,1\}$ with $X \sim \text{Bern}(X|\theta)$

Then,

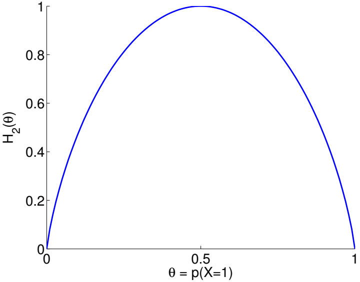
$$p(X = 0) = 1 - \theta$$
$$p(X = 1) = \theta$$

So, the entropy of a Bernoulli random variable is

$$H(X) = -\sum_{x \in \{0,1\}} -p(x) \cdot \log p(x)$$
$$= -\theta \log \theta - (1-\theta)\log(1-\theta)$$

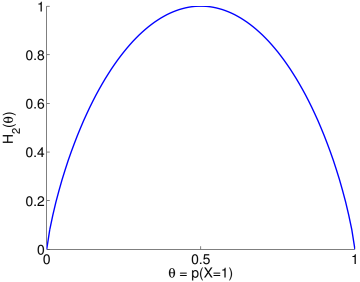
Let $X \in \{0,1\}$ with $X \sim \text{Bern}(X|\theta)$ and $\theta = p(X=1)$

$$H(X) = -\theta \log \theta - (1-\theta) \log(1-\theta)$$



Let $X \in \{0, 1\}$ with $X \sim \text{Bern}(X|\theta)$ and $\theta = p(X = 1)$

$$H(X) = -\theta \log \theta - (1 - \theta) \log(1 - \theta)$$



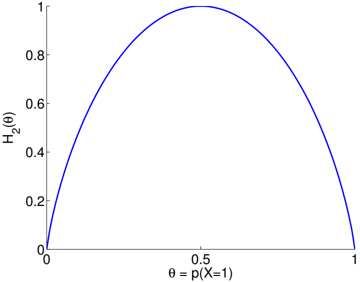
- *Concave* function of the distribution

Entropy of a Random Variable

Example 1 — Bernoulli Distribution

Let $X \in \{0, 1\}$ with $X \sim \text{Bern}(X|\theta)$ and $\theta = p(X = 1)$

$$H(X) = -\theta \log \theta - (1 - \theta) \log(1 - \theta)$$



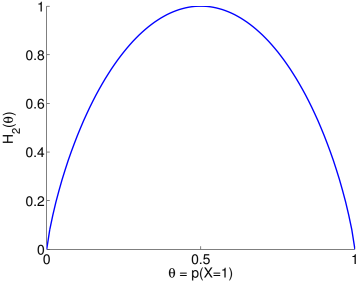
- *Concave* function of the distribution
- Minimum entropy $\rightarrow$ no uncertainty about X, i.e. $\theta = 1$ or $\theta = 0$

Entropy of a Random Variable

Example 1 — Bernoulli Distribution

Let $X \in \{0, 1\}$ with $X \sim \text{Bern}(X|\theta)$ and $\theta = p(X = 1)$

$$H(X) = -\theta \log \theta - (1 - \theta) \log(1 - \theta)$$



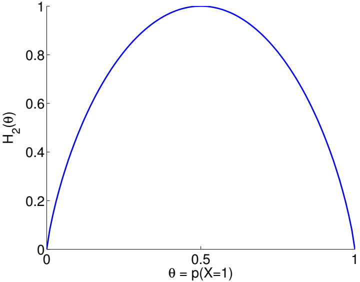
- *Concave* function of the distribution
- Minimum entropy $\rightarrow$ no uncertainty about X, i.e. $\theta = 1$ or $\theta = 0$
- Maximum when $\rightarrow$ complete uncertainty about X, i.e. $\theta = 0.5$

Entropy of a Random Variable

Example 1 — Bernoulli Distribution

Let $X \in \{0, 1\}$ with $X \sim \text{Bern}(X|\theta)$ and $\theta = p(X = 1)$

$$H(X) = -\theta \log \theta - (1-\theta)\log(1-\theta)$$



- *Concave* function of the distribution
- Minimum entropy → no uncertainty about X, i.e. $\theta = 1$ or $\theta = 0$
- Maximum when → complete uncertainty about X, i.e. $\theta = 0.5$
- For $\theta = 0.5$ (e.g. a fair coin) $H_2(X) = 1$ bit.

Consider a random variable X with uniform distribution over 32 outcomes:

The entropy of this rv is given by:

$$H(X) = -\sum_{i=1}^{32} p(i) \log_2 p(i) = -\sum_{i=1}^{32} \frac{1}{32} \log_2 \frac{1}{32} = \log_2 32 = 5 \text{ bits.}$$

Entropy of a Random Variable

Example 3 — Categorical Distribution

Categorical distributions with 30 different states:

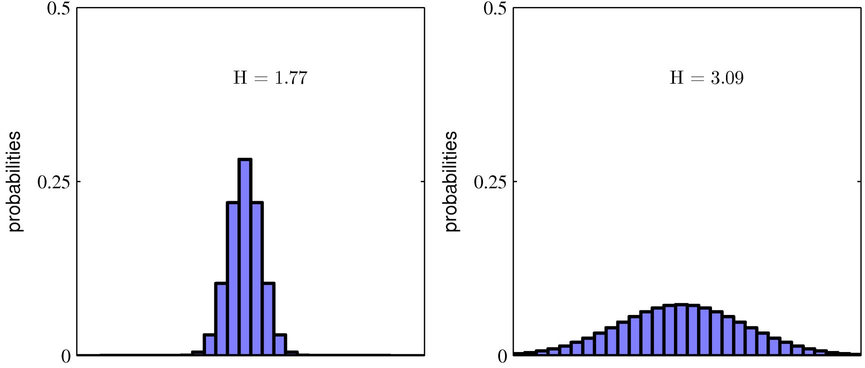


Figure from Bishop, PRML, 2006)

Entropy of a Random Variable

Example 3 — Categorical Distribution

Categorical distributions with 30 different states:

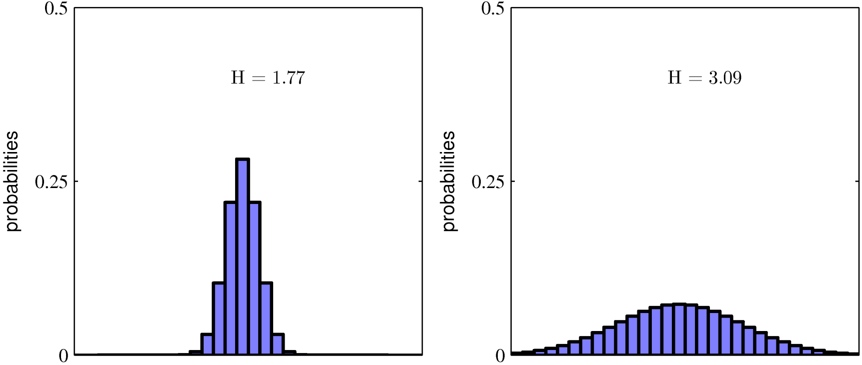


Figure from Bishop, PRML, 2006)

- The more sharply peaked the lower the entropy

Entropy of a Random Variable

Example 3 — Categorical Distribution

Categorical distributions with 30 different states:

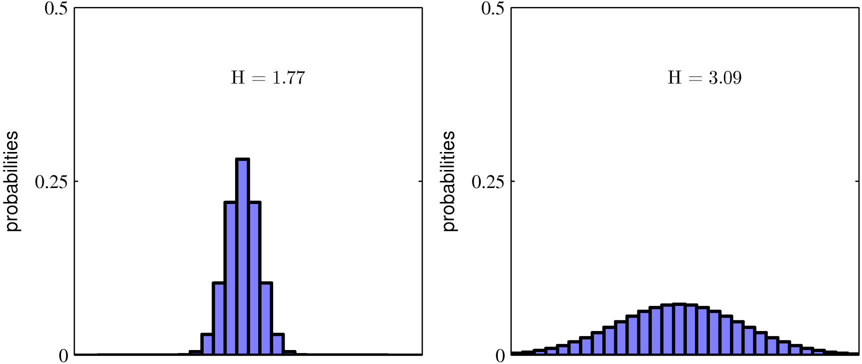


Figure from Bishop, PRML, 2006)

- The more sharply peaked the lower the entropy
- The more evenly spread the higher the entropy

Entropy of a Random Variable

Example 3 — Categorical Distribution

Categorical distributions with 30 different states:

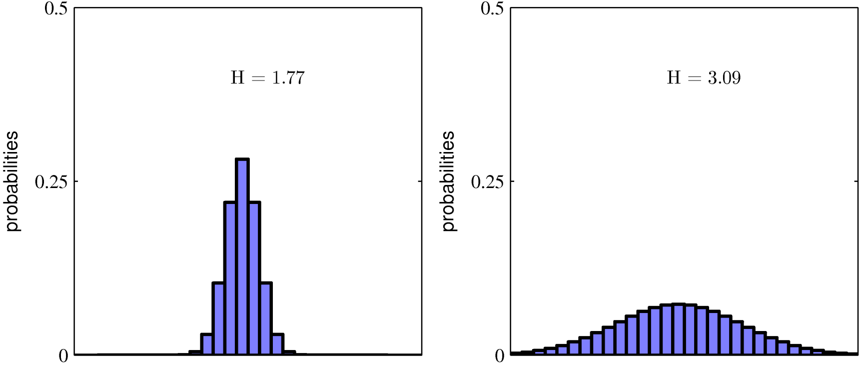


Figure from Bishop, PRML, 2006)

- The more sharply peaked the lower the entropy
- The more evenly spread the higher the entropy
- Maximum for *uniform* distribution: $H(X) = -\log \frac{1}{30} \approx 3.40$ nats
  - When will the entropy be minimum?

Entropy of a Random Variable
Maximum Entropy

Consider a discrete variable X taking on values from the set $\mathcal{X}$

- Let p_i be the probability of each state, with $i = 1, \ldots, |\mathcal{X}|$

Entropy of a Random Variable
Maximum Entropy

Consider a discrete variable X taking on values from the set $\mathcal{X}$

- Let p_i be the probability of each state, with $i = 1, \ldots, |\mathcal{X}|$
- Denote the vector of probabilities with $\mathbf{p}$

Consider a discrete variable X taking on values from the set $\mathcal{X}$

- Let p_i be the probability of each state, with $i = 1, \ldots, |\mathcal{X}|$
- Denote the vector of probabilities with $\mathbf{p}$

The entropy is maximized if $\mathbf{p}$ is uniform:

$$H(X) \leq \log|\mathcal{X}|$$

with equality iff $p_i = \frac{1}{|\mathcal{X}|}$ for all i

Note that $\log|\mathcal{X}|$ is the number of bits needed to describe an outcome of X

Entropy of a Random Variable

Example 4 (from Cover & Thomas, 2006) — 1 of 3

Consider a horse race with 8 horses participating:

$$\{acer, babe, cactus, daisy, epic, fancy, gem, hairy\}$$

Entropy of a Random Variable

Example 4 (from Cover & Thomas, 2006) — 1 of 3

Consider a horse race with 8 horses participating:

$$\{acer, babe, cactus, daisy, epic, fancy, gem, hairy\}$$

- Each horse is equally likely to win. How many bits will we need to transmit the identity of the winning horse?

Entropy of a Random Variable

Example 4 (from Cover & Thomas, 2006) — 1 of 3

Consider a horse race with 8 horses participating:

$$\{acer, babe, cactus, daisy, epic, fancy, gem, hairy\}$$

- Each horse is equally likely to win. How many bits will we need to transmit the identity of the winning horse?

  Note that the entropy of the corresponding random variable, say X is:

$$H(X) = 8 \times \frac{1}{8} \log_2 8 = 3 \text{ bits.}$$

Consider a horse race with 8 horses participating:

$$\{acer, babe, cactus, daisy, epic, fancy, gem, hairy\}$$

- Each horse is equally likely to win. How many bits will we need to transmit the identity of the winning horse?

  Note that the entropy of the corresponding random variable, say X is:

$$H(X) = 8 \times \frac{1}{8} \log_2 8 = 3 \text{ bits.}$$

- Now say that the probabilities of each horse winning are:

$$\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right)$$

Consider a horse race with 8 horses participating:

$$\{acer, babe, cactus, daisy, epic, fancy, gem, hairy\}$$

- Each horse is equally likely to win. How many bits will we need to transmit the identity of the winning horse?

  Note that the entropy of the corresponding random variable, say X is:

$$H(X) = 8 \times \frac{1}{8} \log_2 8 = 3 \text{ bits.}$$

- Now say that the probabilities of each horse winning are:

$$\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right)$$

  What is the average code-length to transmit the identity of the winning horse?

We see that some horses have higher probability of winning:

- We can still use a 3-bit representation
  - However, this would be wasteful as some horses are more likely to win

We see that some horses have higher probability of winning:

- We can still use a 3-bit representation
  - However, this would be wasteful as some horses are more likely to win
- Idea: Use shorter codes for most probable horses and longer codes for the less probable horses.

We see that some horses have higher probability of winning:

- We can still use a 3-bit representation
  - However, this would be wasteful as some horses are more likely to win
- Idea: Use shorter codes for most probable horses and longer codes for the less probable horses.
- Represent the horses (states) using the following codes:

$$\{0, 10, 110, 1110, 111100, 111101, 111110, 111111\}$$

  - We should be able to disambiguate a concatenation of these strings into the corresponding components.
  - E.g. 11001110 $\rightarrow$??

What is the average code length that has to be transmitted?

What is the average code length that has to be transmitted?

$$\text{Average code-length} = \frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{8} \times 3 + \frac{1}{16} \times 4 + 4 \times \frac{1}{64} \times 6 = 2 \text{ bits}$$

What is the average code length that has to be transmitted?

$$\text{Average code-length} = \frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{8} \times 3 + \frac{1}{16} \times 4 + 4 \times \frac{1}{64} \times 6 = 2 \text{ bits}$$

What is the entropy of the corresponding random variable?

What is the average code length that has to be transmitted?

$$\text{Average code-length} = \frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{8} \times 3 + \frac{1}{16} \times 4 + 4 \times \frac{1}{64} \times 6 = 2 \text{ bits}$$

What is the entropy of the corresponding random variable?

$$\begin{aligned} H(X) &= -\left(\frac{1}{2} \log_2 \frac{1}{2} + \frac{1}{4} \log_2 \frac{1}{4} + \frac{1}{8} \log_2 \frac{1}{8} + \frac{1}{16} \log_2 \frac{1}{16} + \frac{4}{64} \log_2 \frac{1}{64} \right) \\ &= 2 \text{ bits} \end{aligned}$$

Entropy of a Random Variable:

Example 5 (from Cover & Thomas, 2006)

Let $X \in \{1, 2, 3\}$ and $p(X = 1) = p(X = 2) = p(X = 3) = \frac{1}{3}$

Given the corresponding codeword:

$$\{\overbrace{0}^{1}, \overbrace{10}^{2}, \overbrace{11}^{3}\}$$

Then $H(X) = 1.58$, and average code length $= 1.66$

Entropy of a Random Variable:
Example 5 (from Cover & Thomas, 2006)

Let $X \in \{1, 2, 3\}$ and $p(X = 1) = p(X = 2) = p(X = 3) = \frac{1}{3}$

Given the corresponding codeword:

$$\{\overbrace{0}^{1}, \overbrace{10}^{2}, \overbrace{11}^{3}\}$$

Then $H(X) = 1.58$, and average code length $= 1.66$

In general, Entropy is a lower bound on the average number of bits to transmit the state of a random variable.

Assume that only the following horses participated in the last race: {*acer*, *babe*, *cactus*, *daisy*}.

The corresponding probabilities of winning are give by:

$$p(X = a) = \frac{1}{2} \quad p(X = b) = \frac{1}{4} \quad p(X = c) = \frac{1}{8} \quad p(X = d) = \frac{1}{8}.$$

You want to determine who won the race with the minimum number of yes/no questions:

(a) What questions should you ask?

(b) What is the minimum expected number of binary questions for this?

Entropy of a Random Variable:

What Questions Should We Ask? (From Cover & Thomas, 2006) — Cont'd

As acer is more likely to have won the race I first ask about him: has $X = a$ won the race?

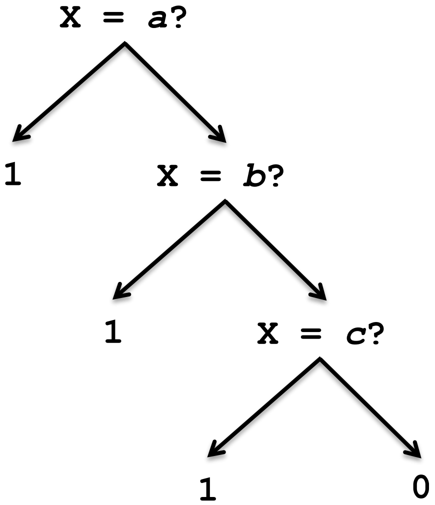
If the answer is no, I then ask about the second most probable winner: has $X = b$ won the race?

Then $X = c$?, and $X = d$?

Note that the series of questions corresponding to an outcome can be seen as a code!

Entropy of a Random Variable:

What Questions Should We Ask? (From Cover & Thomas, 2006) — Cont'd



a	1
b	01
c	001
d	000

The entropy of this random variable determines a lower bound for the minimum number of binary questions:

$$H_2(X) = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{4}\log_2\frac{1}{4} + \frac{1}{8}\log_2\frac{1}{8} + \frac{1}{8}\log_2\frac{1}{8}\right) = 1.75 \text{ bits.}$$

This is in fact the minimum expected number of binary questions. In general, this number lies between $H(X)$ and $H(X) + 1$

Intuitively, each question reduces our amount of uncertainty in the outcome by attempting to eliminate (or validate) the hard to predict outcomes

Joint Entropy

The joint entropy $H(X, Y)$ of a pair of discrete random variables with joint distribution $p(X, Y)$ is given by:

$$
\begin{aligned}
H(X, Y) &= \mathbb{E}_{X,Y}\left[\log \frac{1}{p(X, Y)}\right] \\
&= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{1}{p(x, y)}
\end{aligned}
$$

Joint Entropy:
Independent Random Variables

If X and Y are statistically independent we have that:

$$H(X, Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{1}{p(x, y)}$$

Joint Entropy:
Independent Random Variables

If X and Y are statistically independent we have that:

$$
\begin{aligned}
H(X, Y) &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{1}{p(x, y)} \\
&= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x) p(y) \big[\log p(x) + \log p(y) \big]
\end{aligned}
$$

Joint Entropy:
Independent Random Variables

If X and Y are statistically independent we have that:

$$
\begin{aligned}
H(X,Y) &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{1}{p(x,y)} \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x)p(y)\big[\log p(x) + \log p(y)\big] \\
&= -\sum_{x\in\mathcal{X}} p(x)\log p(x)\underbrace{\sum_{y\in\mathcal{Y}} p(y)}_{1} - \sum_{y\in\mathcal{Y}} p(y)\log p(y)\underbrace{\sum_{x\in\mathcal{X}} p(x)}_{1}
\end{aligned}
$$

Joint Entropy:

Independent Random Variables

If X and Y are statistically independent we have that:

$$
\begin{aligned}
H(X, Y) &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{1}{p(x, y)} \\
&= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x) p(y) \big[\log p(x) + \log p(y) \big] \\
&= -\sum_{x \in \mathcal{X}} p(x) \log p(x) \underbrace{\sum_{y \in \mathcal{Y}} p(y)}_{1} - \sum_{y \in \mathcal{Y}} p(y) \log p(y) \underbrace{\sum_{x \in \mathcal{X}} p(x)}_{1} \\
&= \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)} + \sum_{y \in \mathcal{Y}} p(y) \log \frac{1}{p(y)}
\end{aligned}
$$

Joint Entropy:

Independent Random Variables

If X and Y are statistically independent we have that:

$$
\begin{aligned}
H(X,Y) &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{1}{p(x,y)} \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x)p(y)\big[\log p(x) + \log p(y)\big] \\
&= -\sum_{x\in\mathcal{X}} p(x)\log p(x)\underbrace{\sum_{y\in\mathcal{Y}} p(y)}_{1} - \sum_{y\in\mathcal{Y}} p(y)\log p(y)\underbrace{\sum_{x\in\mathcal{X}} p(x)}_{1} \\
&= \sum_{x\in\mathcal{X}} p(x)\log\frac{1}{p(x)} + \sum_{y\in\mathcal{Y}} p(y)\log\frac{1}{p(y)} \\
&= H(X) + H(Y)
\end{aligned}
$$

Joint Entropy:
Independent Random Variables

If X and Y are statistically independent we have that:

$$
\begin{aligned}
H(X,Y) &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{1}{p(x,y)} \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x)p(y)\left[\log p(x) + \log p(y)\right] \\
&= -\sum_{x\in\mathcal{X}} p(x)\log p(x)\underbrace{\sum_{y\in\mathcal{Y}} p(y)}_{1} - \sum_{y\in\mathcal{Y}} p(y)\log p(y)\underbrace{\sum_{x\in\mathcal{X}} p(x)}_{1} \\
&= \sum_{x\in\mathcal{X}} p(x)\log\frac{1}{p(x)} + \sum_{y\in\mathcal{Y}} p(y)\log\frac{1}{p(y)} \\
&= H(X) + H(Y)
\end{aligned}
$$

Entropy is additive for independent random variables

Conditional Entropy

The conditional entropy of Y given $X = x$ is the entropy of the probability distribution $p(Y|X = x)$:

$$H(Y|X = x) = \sum_{y \in \mathcal{Y}} p(y|X = x) \log \frac{1}{p(y|X = x)}$$

Conditional Entropy

The conditional entropy of Y given $X = x$ is the entropy of the probability distribution $p(Y|X = x)$:

$$H(Y|X = x) = \sum_{y \in \mathcal{Y}} p(y|X = x) \log \frac{1}{p(y|X = x)}$$

The conditional entropy of Y given X, is the average over X of the conditional entropy of Y given $X = x$:

$$\begin{aligned} H(Y|X) &= \sum_{x \in \mathcal{X}} p(x) H(Y|X = x) \\ &= \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log \frac{1}{p(y|x)} \end{aligned}$$

Conditional Entropy

The conditional entropy of Y given $X = x$ is the entropy of the probability distribution $p(Y|X = x)$:

$$H(Y|X = x) = \sum_{y \in \mathcal{Y}} p(y|X = x) \log \frac{1}{p(y|X = x)}$$

The conditional entropy of Y given X, is the average over X of the conditional entropy of Y given $X = x$:

$$\begin{aligned} H(Y|X) &= \sum_{x \in \mathcal{X}} p(x) H(Y|X = x) \\ &= \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log \frac{1}{p(y|x)} \end{aligned}$$

Average uncertainty that remains about Y when X is known.

Conditional Entropy — Cont'd

We can re-write the conditional entropy as follows:

$$H(Y|X) = \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log \frac{1}{p(y|x)}$$

Conditional Entropy — Cont'd

We can re-write the conditional entropy as follows:

$$
\begin{aligned}
H(Y|X) &= \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log \frac{1}{p(y|x)} \\
&= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x) p(y|x) \log \frac{1}{p(y|x)}
\end{aligned}
$$

Conditional Entropy — Cont'd

We can re-write the conditional entropy as follows:

$$
\begin{aligned}
H(Y|X) &= \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log \frac{1}{p(y|x)} \\
&= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x) p(y|x) \log \frac{1}{p(y|x)} \\
&= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{1}{p(y|x)}
\end{aligned}
$$

Conditional Entropy — Cont'd

We can re-write the conditional entropy as follows:

$$
\begin{aligned}
H(Y|X) &= \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log \frac{1}{p(y|x)} \\
&= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x) p(y|x) \log \frac{1}{p(y|x)} \\
&= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{1}{p(y|x)} \\
&= \mathbb{E}_{X,Y} \left[\log \frac{1}{p(Y|X)} \right]
\end{aligned}
$$

Conditional Entropy — Cont'd

We can re-write the conditional entropy as follows:

$$
\begin{aligned}
H(Y|X) &= \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log \frac{1}{p(y|x)} \\
&= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x) p(y|x) \log \frac{1}{p(y|x)} \\
&= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{1}{p(y|x)} \\
&= \mathbb{E}_{X,Y} \left[\log \frac{1}{p(Y|X)} \right]
\end{aligned}
$$

Note the expectation is not wrt the conditional distribution but wrt the joint distribution $p(X, Y)$

Chain Rule

The joint entropy can be written as:

$$H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y)$$

Chain Rule

The joint entropy can be written as:

$$\begin{aligned} H(X,Y) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y) \\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\big[\log p(x) + \log p(y|x)\big] \end{aligned}$$

Chain Rule

The joint entropy can be written as:

$$
\begin{aligned}
H(X,Y) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y) \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\big[\log p(x) + \log p(y|x)\big] \\
&= -\sum_{x\in\mathcal{X}} \log p(x) \underbrace{\sum_{y\in\mathcal{Y}} p(x,y)}_{p(x)} - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x)
\end{aligned}
$$

Chain Rule

The joint entropy can be written as:

$$\begin{aligned} H(X,Y) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y) \\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\big[\log p(x) + \log p(y|x)\big] \\ &= -\sum_{x\in\mathcal{X}} \log p(x) \underbrace{\sum_{y\in\mathcal{Y}} p(x,y)}_{p(x)} - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x) \end{aligned}$$

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$

Chain Rule

The joint entropy can be written as:

$$
\begin{aligned}
H(X, Y) &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y) \\
&= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \big[\log p(x) + \log p(y|x) \big] \\
&= -\sum_{x \in \mathcal{X}} \log p(x) \underbrace{\sum_{y \in \mathcal{Y}} p(x, y)}_{p(x)} - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y|x)
\end{aligned}
$$

$$H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$

The joint uncertainty of X and Y is the uncertainty of X plus the uncertainty of Y given X

An Axiomatic Characterisation

Suppose we want a measure H of "information" in a random variable X such that

1. H depends on the distribution of X, and not the outcomes themselves
2. The H for the combination of two variables X, Y is at most the sum of the corresponding H values
3. The H for the combination of two independent variables X, Y is the sum of the corresponding H values
4. Adding outcomes with probability zero does not affect H
5. The H for an unbiased Bernoulli is 1
6. The H for a Bernoulli with parameter p tends to 0 as $p \to 0$

Then, the only possible choice for H is

$$H(X) = -\sum_{x} p(x) \log_2 p(x)$$

Outline

Summary

- Entropy as a measure of information content
- Computation of entropy of discrete random variables
- Entropy and average code length
- Entropy and minimum expected number of binary questions
- Joint and conditional entropies, chain rule
- Reading: Mackay § 1.2 – § 1.5, § 8.1; Cover & Thomas § 2.1; Bishop § 1.6

Next time

- More properties of entropy
- Relative entropy
- Mutual information