Outline

Entropy

Noiseless coding theory

Conditional entropy
Entropy

Noiseless coding theory

Conditional entropy
What is entropy?

- Entropy is an important notion in thermodynamics, information theory, data compression, cryptography, etc.
What is entropy?

- Entropy is an important notion in thermodynamics, information theory, data compression, cryptography, etc.
- Familiar on some level to everyone who has studied chemistry or statistical physics.
What is entropy?

- Entropy is an important notion in thermodynamics, information theory, data compression, cryptography, etc.
- Familiar on some level to everyone who has studied chemistry or statistical physics.
- Kind of means amount of randomness or disorder.
What is entropy?

- Entropy is an important notion in thermodynamics, information theory, data compression, cryptography, etc.
- Familiar on some level to everyone who has studied chemistry or statistical physics.
- Kind of means amount of randomness or disorder.
- But can we give a mathematical definition? In particular, how do we define the entropy of a random variable?
Information

- Suppose we toss a fair coin $k$ times.
• Suppose we toss a fair coin \( k \) times.
• Then the state space \( S \) is the set of \( 2^k \) possible heads-tails sequences.
Suppose we toss a fair coin \( k \) times.

Then the state space \( S \) is the set of \( 2^k \) possible heads-tails sequences.

If \( X \) is the random sequence (so \( X \) is a random variable), then for each \( x \in S \) we have \( P\{X = x\} = 2^{-k} \).
Suppose we toss a fair coin \( k \) times.

Then the state space \( S \) is the set of \( 2^k \) possible heads-tails sequences.

If \( X \) is the random sequence (so \( X \) is a random variable), then for each \( x \in S \) we have \( P\{X = x\} = 2^{-k} \).

In information theory it’s quite common to use log to mean \( \log_2 \) instead of \( \log_e \). We follow that convention in this lecture. In particular, this means that

\[
\log P\{X = x\} = -k
\]

for each \( x \in S \).
Suppose we toss a fair coin $k$ times.

Then the state space $S$ is the set of $2^k$ possible heads-tails sequences.

If $X$ is the random sequence (so $X$ is a random variable), then for each $x \in S$ we have $P\{X = x\} = 2^{-k}$.

In information theory it’s quite common to use log to mean $\log_2$ instead of $\log_e$. We follow that convention in this lecture. In particular, this means that

$$\log P\{X = x\} = -k$$

for each $x \in S$.

Since there are $2^k$ values in $S$, it takes $k$ “bits” to describe an element $x \in S$. 
Suppose we toss a fair coin $k$ times.

Then the state space $S$ is the set of $2^k$ possible heads-tails sequences.

If $X$ is the random sequence (so $X$ is a random variable), then for each $x \in S$ we have $P\{X = x\} = 2^{-k}$.

In information theory it’s quite common to use log to mean $\log_2$ instead of $\log_e$. We follow that convention in this lecture. In particular, this means that

$$\log P\{X = x\} = -k$$

for each $x \in S$.

Since there are $2^k$ values in $S$, it takes $k$ “bits” to describe an element $x \in S$.

Intuitively, could say that when we learn that $X = x$, we have learned $k = - \log P\{X = x\}$ “bits of information”.
Shannon entropy


- Goal is to define a notion of how much we “expect to learn” from a random variable or “how many bits of information a random variable contains” that makes sense for general experiments (which may not have anything to do with coins).

- If a random variable $X$ takes values $x_1, x_2, \ldots, x_n$ with positive probabilities $p_1, p_2, \ldots, p_n$ then we define the entropy of $X$ by

$$H(X) = \sum_{i=1}^{n} p_i \log p_i = -\sum_{i=1}^{n} p_i \log p_i.$$
Shannon entropy

- Goal is to define a notion of how much we “expect to learn” from a random variable or “how many bits of information a random variable contains” that makes sense for general experiments (which may not have anything to do with coins).
Shannon entropy

- Goal is to define a notion of how much we “expect to learn” from a random variable or “how many bits of information a random variable contains” that makes sense for general experiments (which may not have anything to do with coins).
- If a random variable $X$ takes values $x_1, x_2, \ldots, x_n$ with positive probabilities $p_1, p_2, \ldots, p_n$ then we define the entropy of $X$ by

$$H(X) = \sum_{i=1}^{n} p_i (-\log p_i) = -\sum_{i=1}^{n} p_i \log p_i.$$
Shannon entropy

- Goal is to define a notion of how much we “expect to learn” from a random variable or “how many bits of information a random variable contains” that makes sense for general experiments (which may not have anything to do with coins).
- If a random variable $X$ takes values $x_1, x_2, \ldots, x_n$ with positive probabilities $p_1, p_2, \ldots, p_n$ then we define the entropy of $X$ by

  \[ H(X) = \sum_{i=1}^{n} p_i (-\log p_i) = -\sum_{i=1}^{n} p_i \log p_i. \]

- This can be interpreted as the expectation of $(-\log p_i)$. The value $(-\log p_i)$ is the “amount of surprise” when we see $x_i$. 
## Twenty questions with Harry

Harry always thinks of one of the following animals:

<table>
<thead>
<tr>
<th>$x$</th>
<th>$P{X = x}$</th>
<th>$- \log P{X = x}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dog</td>
<td>1/4</td>
<td>2</td>
</tr>
<tr>
<td>Cat</td>
<td>1/4</td>
<td>2</td>
</tr>
<tr>
<td>Cow</td>
<td>1/8</td>
<td>3</td>
</tr>
<tr>
<td>Pig</td>
<td>1/16</td>
<td>4</td>
</tr>
<tr>
<td>Squirrel</td>
<td>1/16</td>
<td>4</td>
</tr>
<tr>
<td>Mouse</td>
<td>1/16</td>
<td>4</td>
</tr>
<tr>
<td>Owl</td>
<td>1/16</td>
<td>4</td>
</tr>
<tr>
<td>Sloth</td>
<td>1/32</td>
<td>5</td>
</tr>
<tr>
<td>Hippo</td>
<td>1/32</td>
<td>5</td>
</tr>
<tr>
<td>Yak</td>
<td>1/32</td>
<td>5</td>
</tr>
<tr>
<td>Zebra</td>
<td>1/64</td>
<td>6</td>
</tr>
<tr>
<td>Rhino</td>
<td>1/64</td>
<td>6</td>
</tr>
</tbody>
</table>

Can learn animal with $H(x)$ questions on average.

**General:** expect $H(x)$ questions if probabilities powers of 2. Otherwise $H(x) + 1$ suffice. (Try rounding down to 2 powers.)
Twenty questions with Harry

Harry always thinks of one of the following animals:

<table>
<thead>
<tr>
<th>x</th>
<th>$P{X = x}$</th>
<th>$-\log P{X = x}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dog</td>
<td>1/4</td>
<td>2</td>
</tr>
<tr>
<td>Cat</td>
<td>1/4</td>
<td>2</td>
</tr>
<tr>
<td>Cow</td>
<td>1/8</td>
<td>3</td>
</tr>
<tr>
<td>Pig</td>
<td>1/16</td>
<td>4</td>
</tr>
<tr>
<td>Squirrel</td>
<td>1/16</td>
<td>4</td>
</tr>
<tr>
<td>Mouse</td>
<td>1/16</td>
<td>4</td>
</tr>
<tr>
<td>Owl</td>
<td>1/16</td>
<td>4</td>
</tr>
<tr>
<td>Sloth</td>
<td>1/32</td>
<td>5</td>
</tr>
<tr>
<td>Hippo</td>
<td>1/32</td>
<td>5</td>
</tr>
<tr>
<td>Yak</td>
<td>1/32</td>
<td>5</td>
</tr>
<tr>
<td>Zebra</td>
<td>1/64</td>
<td>6</td>
</tr>
<tr>
<td>Rhino</td>
<td>1/64</td>
<td>6</td>
</tr>
</tbody>
</table>

Can learn animal with $H(X)$ questions on average.
Twenty questions with Harry

- Harry always thinks of one of the following animals:

<table>
<thead>
<tr>
<th>x</th>
<th>$P{X = x}$</th>
<th>$-\log P{X = x}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dog</td>
<td>1/4</td>
<td>2</td>
</tr>
<tr>
<td>Cat</td>
<td>1/4</td>
<td>2</td>
</tr>
<tr>
<td>Cow</td>
<td>1/8</td>
<td>3</td>
</tr>
<tr>
<td>Pig</td>
<td>1/16</td>
<td>4</td>
</tr>
<tr>
<td>Squirrel</td>
<td>1/16</td>
<td>4</td>
</tr>
<tr>
<td>Mouse</td>
<td>1/16</td>
<td>4</td>
</tr>
<tr>
<td>Owl</td>
<td>1/16</td>
<td>4</td>
</tr>
<tr>
<td>Sloth</td>
<td>1/32</td>
<td>5</td>
</tr>
<tr>
<td>Hippo</td>
<td>1/32</td>
<td>5</td>
</tr>
<tr>
<td>Yak</td>
<td>1/32</td>
<td>5</td>
</tr>
<tr>
<td>Zebra</td>
<td>1/64</td>
<td>6</td>
</tr>
<tr>
<td>Rhino</td>
<td>1/64</td>
<td>6</td>
</tr>
</tbody>
</table>

- Can learn animal with $H(X)$ questions on average.

- **General:** expect $H(X)$ questions if probabilities powers of 2. Otherwise $H(X) + 1$ suffice. (Try rounding down to 2 powers.)
Again, if a random variable $X$ takes the values $x_1, x_2, \ldots, x_n$ with positive probabilities $p_1, p_2, \ldots, p_n$ then we define the **entropy** of $X$ by

$$H(X) = \sum_{i=1}^{n} p_i (-\log p_i) = -\sum_{i=1}^{n} p_i \log p_i.$$
Other examples

- Again, if a random variable $X$ takes the values $x_1, x_2, \ldots, x_n$ with positive probabilities $p_1, p_2, \ldots, p_n$ then we define the entropy of $X$ by

$$H(X) = \sum_{i=1}^{n} p_i (- \log p_i) = - \sum_{i=1}^{n} p_i \log p_i.$$ 

- If $X$ takes one value with probability 1, what is $H(X)$?
Again, if a random variable $X$ takes the values $x_1, x_2, \ldots, x_n$ with positive probabilities $p_1, p_2, \ldots, p_n$ then we define the **entropy** of $X$ by

$$H(X) = \sum_{i=1}^{n} p_i (\log p_i) = -\sum_{i=1}^{n} p_i \log p_i.$$ 

- If $X$ takes one value with probability 1, what is $H(X)$?
- If $X$ takes $k$ values with equal probability, what is $H(X)$?
Again, if a random variable $X$ takes the values $x_1, x_2, \ldots, x_n$ with positive probabilities $p_1, p_2, \ldots, p_n$ then we define the **entropy** of $X$ by

$$H(X) = \sum_{i=1}^{n} p_i (-\log p_i) = -\sum_{i=1}^{n} p_i \log p_i.$$ 

- If $X$ takes one value with probability 1, what is $H(X)$?
- If $X$ takes $k$ values with equal probability, what is $H(X)$?
- What is $H(X)$ if $X$ is a geometric random variable with parameter $p = 1/2$?
Consider random variables $X, Y$ with joint mass function $p(x_i, y_j) = P\{X = x_i, Y = y_j\}$. Then we write

$$H(X, Y) = -\sum_{i} \sum_{j} p(x_i, y_j) \log p(x_i, y_j).$$

$H(X, Y)$ is just the entropy of the pair $(X, Y)$ (viewed as a random variable itself).

Claim: if $X$ and $Y$ are independent, then $H(X, Y) = H(X) + H(Y)$. Why is that?
Consider random variables $X, Y$ with joint mass function $p(x_i, y_j) = P\{X = x_i, Y = y_j\}$.

Then we write

$$H(X, Y) = - \sum_i \sum_j p(x_i, y_j) \log p(x_i, y_j).$$
Consider random variables $X, Y$ with joint mass function 
$$p(x_i, y_j) = P\{X = x_i, Y = y_j\}.$$  
Then we write 
$$H(X, Y) = -\sum_i \sum_j p(x_i, y_j) \log p(x_i, y_j).$$  
$H(X, Y)$ is just the entropy of the pair $(X, Y)$ (viewed as a random variable itself).
Entropy for a pair of random variables

Consider random variables $X, Y$ with joint mass function $p(x_i, y_j) = P\{X = x_i, Y = y_j\}$.

Then we write

\[
H(X, Y) = - \sum_i \sum_j p(x_i, y_j) \log p(x_i, y_j).
\]

$H(X, Y)$ is just the entropy of the pair $(X, Y)$ (viewed as a random variable itself).

Claim: if $X$ and $Y$ are independent, then

\[
H(X, Y) = H(X) + H(Y).
\]

Why is that?
Outline

Entropy

Noiseless coding theory

Conditional entropy
Outline

Entropy

Noiseless coding theory

Conditional entropy
Coding values by bit sequences

- If $X$ takes four values $A, B, C, D$ we can code them by:
  
  $A \leftrightarrow 00$
  
  $B \leftrightarrow 01$
  
  $C \leftrightarrow 10$
  
  $D \leftrightarrow 11$

If $X$ takes four values $A, B, C, D$ we can code them by:

- $A \leftrightarrow 00$
- $B \leftrightarrow 01$
- $C \leftrightarrow 10$
- $D \leftrightarrow 11$

Or by

- $A \leftrightarrow 0$
- $B \leftrightarrow 10$
- $C \leftrightarrow 110$
- $D \leftrightarrow 111$

What does 100111110010 spell?

A coding scheme is equivalent to a twenty questions strategy.
Coding values by bit sequences


If $X$ takes four values $A, B, C, D$ we can code them by:

\[
A \leftrightarrow 00 \\
B \leftrightarrow 01 \\
C \leftrightarrow 10 \\
D \leftrightarrow 11
\]

Or by

\[
A \leftrightarrow 0 \\
B \leftrightarrow 10 \\
C \leftrightarrow 110 \\
D \leftrightarrow 111
\]

No sequence in code is an extension of another.

If $X$ takes four values $A, B, C, D$ we can code them by:

- $A \leftrightarrow 00$
- $B \leftrightarrow 01$
- $C \leftrightarrow 10$
- $D \leftrightarrow 11$

Or by:

- $A \leftrightarrow 0$
- $B \leftrightarrow 10$
- $C \leftrightarrow 110$
- $D \leftrightarrow 111$

No sequence in code is an extension of another.

What does 100111110010 spell?
Coding values by bit sequences

- If $X$ takes four values $A, B, C, D$ we can code them by:
  
  \[
  \begin{align*}
  A & \leftrightarrow 00 \\
  B & \leftrightarrow 01 \\
  C & \leftrightarrow 10 \\
  D & \leftrightarrow 11
  \end{align*}
  \]

- Or by
  
  \[
  \begin{align*}
  A & \leftrightarrow 0 \\
  B & \leftrightarrow 10 \\
  C & \leftrightarrow 110 \\
  D & \leftrightarrow 111
  \end{align*}
  \]

- No sequence in code is an extension of another.
- What does $100111110010$ spell?
- A coding scheme is equivalent to a twenty questions strategy.
Noiseless coding theorem: Expected number of questions you need is always at least the entropy.
Twenty questions theorem

- **Noiseless coding theorem:** Expected number of questions you need is always at least the entropy.
- **Note:** The expected number of questions is the entropy if each question divides the space of possibilities exactly in half (measured by probability).
Noiseless coding theorem: Expected number of questions you need is always at least the entropy.

Note: The expected number of questions is the entropy if each question divides the space of possibilities exactly in half (measured by probability).

In this case, let $X$ take values $x_1, \ldots, x_N$ with probabilities $p(x_1), \ldots, p(x_N)$. Then if a valid coding of $X$ assigns $n_i$ bits to $x_i$, we have

$$\sum_{i=1}^{N} n_i p(x_i) \geq H(X) = - \sum_{i=1}^{N} p(x_i) \log p(x_i).$$
Noiseless coding theorem: Expected number of questions you need is always at least the entropy.

Note: The expected number of questions is the entropy if each question divides the space of possibilities exactly in half (measured by probability).

In this case, let $X$ take values $x_1,\ldots,x_N$ with probabilities $p(x_1),\ldots,p(x_N)$. Then if a valid coding of $X$ assigns $n_i$ bits to $x_i$, we have

$$\sum_{i=1}^{N} n_i p(x_i) \geq H(X) = -\sum_{i=1}^{N} p(x_i) \log p(x_i).$$

Data compression: $X_1, X_2,\ldots,X_n$ be i.i.d. instances of $X$. Do there exist encoding schemes such that the expected number of bits required to encode the entire sequence is about $H(X)n$ (assuming $n$ is sufficiently large)?
Twenty questions theorem

- **Noiseless coding theorem:** Expected number of questions you need is always at least the entropy.
- **Note:** The expected number of questions is the entropy if each question divides the space of possibilities exactly in half (measured by probability).
- In this case, let $X$ take values $x_1, \ldots, x_N$ with probabilities $p(x_1), \ldots, p(x_N)$. Then if a valid coding of $X$ assigns $n_i$ bits to $x_i$, we have

$$
\sum_{i=1}^{N} n_i p(x_i) \geq H(X) = - \sum_{i=1}^{N} p(x_i) \log p(x_i).
$$

- **Data compression:** $X_1, X_2, \ldots, X_n$ be i.i.d. instances of $X$. Do there exist encoding schemes such that the expected number of bits required to encode the entire sequence is about $H(X)n$ (assuming $n$ is sufficiently large)?
- Yes. Consider space of $N^n$ possibilities. Use “rounding to 2 power” trick, Expect to need at most $H(x)n + 1$ bits.
Entropy

Noiseless coding theory

Conditional entropy
Outline

Entropy

Noiseless coding theory

Conditional entropy
Let’s again consider random variables $X, Y$ with joint mass function $p(x_i, y_j) = P\{X = x_i, Y = y_j\}$ and write

$$H(X, Y) = - \sum_i \sum_j p(x_i, y_j) \log p(x_i, y_j).$$
Conditional entropy

Let’s again consider random variables $X, Y$ with joint mass function $p(x_i, y_j) = P\{X = x_i, Y = y_j\}$ and write

$$H(X, Y) = -\sum_i \sum_j p(x_i, y_j) \log p(x_i, y_i).$$

But now let’s not assume they are independent.
Conditional entropy

Let’s again consider random variables $X, Y$ with joint mass function $p(x_i, y_j) = P\{X = x_i, Y = y_j\}$ and write

$$H(X, Y) = - \sum_i \sum_j p(x_i, y_j) \log p(x_i, y_j).$$

But now let’s not assume they are independent.

We can define a **conditional entropy** of $X$ given $Y = y_j$ by

$$H_{Y=y_j}(X) = - \sum_i p(x_i|y_j) \log p(x_i|y_j).$$
Let’s again consider random variables $X, Y$ with joint mass function $p(x_i, y_j) = P\{X = x_i, Y = y_j\}$ and write

$$H(X, Y) = -\sum_i \sum_j p(x_i, y_j) \log p(x_i, y_j).$$

But now let’s not assume they are independent.

We can define a **conditional entropy** of $X$ given $Y = y_j$ by

$$H_{Y=y_j}(X) = -\sum_i p(x_i|y_j) \log p(x_i|y_j).$$

This is just the entropy of the conditional distribution. Recall that $p(x_i|y_j) = P\{X = x_i|Y = y_j\}$. 

Conditional entropy

- Let’s again consider random variables $X, Y$ with joint mass function $p(x_i, y_j) = P\{X = x_i, Y = y_j\}$ and write

$$H(X, Y) = -\sum_i \sum_j p(x_i, y_j) \log p(x_i, y_j).$$

- But now let’s not assume they are independent.

- We can define a **conditional entropy** of $X$ given $Y = y_j$ by

$$H_{Y=y_j}(X) = -\sum_i p(x_i|y_j) \log p(x_i|y_j).$$

- This is just the entropy of the conditional distribution. Recall that $p(x_i|y_j) = P\{X = x_i|Y = y_j\}$.

- We similarly define $H_Y(X) = \sum_j H_{Y=y_j}(X)p_Y(y_j)$. This is the *expected* amount of conditional entropy that there will be in $Y$ after we have observed $X$. 
Properties of conditional entropy

Definitions: \( H_{Y=y_j}(X) = -\sum_i p(x_i|y_j) \log p(x_i|y_j) \) and \( H_Y(X) = \sum_j H_{Y=y_j}(X) p_Y(y_j) \).

Important property one:
\[
H(X, Y) = H(Y) + H_{Y|(X)}
\]

In words, the expected amount of information we learn when discovering \((X, Y)\) is equal to expected amount we learn when discovering \(Y\) plus expected amount when we subsequently discover \(X\) (given our knowledge of \(Y\)).

To prove this property, recall that \( p(x_i, y_j) = p_Y(y_j) p(x_i|y_j) \).

Thus,
\[
H(X, Y) = -\sum_i \sum_j p(x_i, y_j) \log p(x_i, y_j) = -\sum_i \sum_j p(y_j) p(x_i|y_j) \log p(x_i|y_j) = -\sum_j p_Y(y_j) \sum_i p(x_i|y_j) \log p(x_i|y_j) = -\sum_j p_Y(y_j) \sum_i p(x_i|y_j) \log p(x_i|y_j) = \sum_j H_{Y=y_j}(X) p_Y(y_j).
\]
Properties of conditional entropy

- Definitions: $H_{Y=y_j}(X) = - \sum_i p(x_i|y_j) \log p(x_i|y_j)$ and $H_Y(X) = \sum_j H_{Y=y_j}(X) p_Y(y_j)$.
- Important property one: $H(X, Y) = H(Y) + H_Y(X)$. 

Definitions: $H_{Y=y_j}(X) = -\sum_i p(x_i|y_j) \log p(x_i|y_j)$ and $H_Y(X) = \sum_j H_{Y=y_j}(X)p_Y(y_j)$.

**Important property one:** $H(X, Y) = H(Y) + H_Y(X)$.

In words, the expected amount of information we learn when discovering $(X, Y)$ is equal to expected amount we learn when discovering $Y$ plus expected amount when we subsequently discover $X$ (given our knowledge of $Y$).
Properties of conditional entropy

- **Definitions:** \( H_{Y=y_j}(X) = - \sum_i p(x_i|y_j) \log p(x_i|y_j) \) and \( H_Y(X) = \sum_j H_{Y=y_j}(X)p_Y(y_j) \).

- **Important property one:** \( H(X, Y) = H(Y) + H_Y(X) \).

- In words, the expected amount of information we learn when discovering \((X, Y)\) is equal to expected amount we learn when discovering \(Y\) plus expected amount when we subsequently discover \(X\) (given our knowledge of \(Y\)).

- To prove this property, recall that \( p(x_i, y_j) = p_Y(y_j)p(x_i|y_j) \).
Properties of conditional entropy

Definitions: \( H_{Y=y_j}(X) = - \sum_i p(x_i|y_j) \log p(x_i|y_j) \) and \( H_Y(X) = \sum_j H_{Y=y_j}(X)p_Y(y_j) \).

Important property one: \( H(X, Y) = H(Y) + H_Y(X) \).

In words, the expected amount of information we learn when discovering \((X, Y)\) is equal to expected amount we learn when discovering \(Y\) plus expected amount when we subsequently discover \(X\) (given our knowledge of \(Y\)).

To prove this property, recall that \( p(x_i, y_j) = p_Y(y_j)p(x_i|y_j) \).

Thus, \( H(X, Y) = - \sum_i \sum_j p(x_i, y_j) \log p(x_i, y_j) = - \sum_i \sum_j p_Y(y_j)p(x_i|y_j)[\log p_Y(y_j) + \log p(x_i|y_j)] = - \sum_j p_Y(y_j) \log p_Y(y_j) \sum_i p(x_i|y_j) - \sum_j p_Y(y_j) \sum_i p(x_i|y_j) \log p(x_i|y_j) = H(Y) + H_Y(X) \).
Properties of conditional entropy

- Definitions: \( H_{Y=y_j}(X) = - \sum_i p(x_i|y_j) \log p(x_i|y_j) \) and \( H_Y(X) = \sum_j H_{Y=y_j}(X)p_Y(y_j) \).

- Important property two: \( H_Y(X) \leq H(X) \) with equality if and only if \( X \) and \( Y \) are independent.

- In words, the expected amount of information we learn when discovering \( X \) after having discovered \( Y \) can't be more than the expected amount of information we would learn when discovering \( X \) before knowing anything about \( Y \).
Properties of conditional entropy

Definitions: $H_{Y=y_j}(X) = - \sum_i p(x_i|y_j) \log p(x_i|y_j)$ and $H_Y(X) = \sum_j H_{Y=y_j}(X)p_Y(y_j)$.

Important property two: $H_Y(X) \leq H(X)$ with equality if and only if $X$ and $Y$ are independent.
Properties of conditional entropy

- **Definitions:** \( H_{Y=y_j}(X) = -\sum_i p(x_i|y_j) \log p(x_i|y_j) \) and \( H_Y(X) = \sum_j H_{Y=y_j}(X)p_Y(y_j) \).
- **Important property two:** \( H_Y(X) \leq H(X) \) with equality if and only if \( X \) and \( Y \) are independent.
- In words, the expected amount of information we learn when discovering \( X \) *after* having discovered \( Y \) can’t be more than the expected amount of information we would learn when discovering \( X \) *before* knowing anything about \( Y \).
Properties of conditional entropy

Definitions: \( H_{Y=y_j}(X) = -\sum_i p(x_i|y_j) \log p(x_i|y_j) \) and \( H_Y(X) = \sum_j H_{Y=y_j}(X)p_Y(y_j) \).

Important property two: \( H_Y(X) \leq H(X) \) with equality if and only if \( X \) and \( Y \) are independent.

In words, the expected amount of information we learn when discovering \( X \) after having discovered \( Y \) can’t be more than the expected amount of information we would learn when discovering \( X \) before knowing anything about \( Y \).

Proof: note that \( \mathcal{E}(p_1, p_2, \ldots, p_n) := -\sum_i p_i \log p_i \) is concave.
Properties of conditional entropy

Definitions: \( H_{Y=y_j}(X) = - \sum_i p(x_i|y_j) \log p(x_i|y_j) \) and \( H_Y(X) = \sum_j H_{Y=y_j}(X)p_Y(y_j) \).

Important property two: \( H_Y(X) \leq H(X) \) with equality if and only if \( X \) and \( Y \) are independent.

In words, the expected amount of information we learn when discovering \( X \) after having discovered \( Y \) can’t be more than the expected amount of information we would learn when discovering \( X \) before knowing anything about \( Y \).

Proof: note that \( \mathcal{E}(p_1, p_2, \ldots, p_n) := - \sum p_i \log p_i \) is concave.

The vector \( v = \{p_X(x_1), p_X(x_2), \ldots, p_X(x_n)\} \) is a weighted average of vectors \( v_j := \{p_X(x_1|y_j), p_X(x_2|y_j), \ldots, p_X(x_n|y_j)\} \) as \( j \) ranges over possible values. By (vector version of) Jensen’s inequality, \( H(X) = \mathcal{E}(v) = \mathcal{E}(\sum p_Y(y_j)v_j) \geq \sum p_Y(y_j)\mathcal{E}(v_j) = H_Y(X) \).