Dependent Random Variables

Mário S. Alvim
(msalvim@dcc.ufmg.br)

Information Theory

DCC-UFMG
(2017/02)

Dependent random variables - Introduction

- In previous lectures we studied measures of the information contents of individual ensembles:
  - the information content $h(x)$ of an event x, and
  - the Shannon entropy $H(X)$ of a single ensemble X.
- In this lecture we will study the information content and entropy of **dependent ensembles** (or random variables).

  In particular, we will define the concepts of **conditional entropy** and of **mutual information**, and we will study their mathematical properties.
- These concepts will be later used in a variety of scenarios: from communication theory to machine learning and artificial intelligence, and from biology to security and privacy.

Dependent random variables - Introduction

- As a motivation, we note that many problems can be modeled more or less as follows:

  1. We are interested in learning the outcome of an ensemble X.
  2. We don't have direct access to X. Instead, we have access to the outcomes of another ensemble Y that may be correlated to X.
  3. We want to infer as much information as possible about X only using what we observed from Y.

  In this kind of problem we need to define and measure the amount of information that an ensemble carries about another ensemble.

Dependent random variables - Introduction

- For instance, to reliably send a message through a communication channel we must know how the output from the channel (i.e., what we can actually see) is correlated with the input to the channel (i.e., what we can't see but want to infer).

  Input and output to a channel can be modeled as dependent ensembles X and Y, respectively, and by measuring the amount of information Y carries about X we can quantify how good the channel is in transmitting information.

Review of Shannon information content and Entropy

Shannon information content

- The **Shannon information content of an outcome** x is defined as

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

  and it is measured in **bits**.

- $h(x)$ measures the surprise one has when learning that the outcome of a random variable was x.

- **Additivity in case of independence:** If x, y are independent, then $h(x, y) = h(x) + h(y)$.

Entropy

- The **entropy of an ensemble** X is defined as

$$H(X) = \sum_{x \in \mathcal{A}_X} p(x) \log_2 \frac{1}{p(x)},$$

with the convention that for $p(x) = 0$ we use $0 \cdot \log_2 1/0 = 0$.

- The entropy of an ensemble is the expected value of the information content of each of its outcomes:

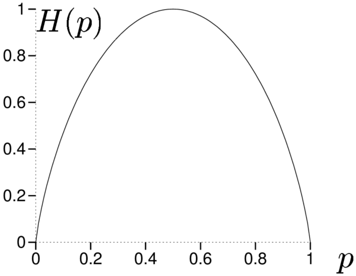
$$H(X) = \sum_{x \in \mathcal{A}_X} p(x) h(x).$$

- If a probability distribution has only two values p and $1 - p$, we define the entropy

$$H(p) = p \log_2 \frac{1}{p} + (1 - p) \log_2 \frac{1}{1 - p}.$$

Entropy - Properties

- **Bounds:** $0 \leq H(X) \leq \log_2 |\mathcal{A}_X|$.
- **Changing the base of the entropy:** $H_b(X) = (\log_a b) \cdot H_a(X)$.
- **Concavity:** $H(p)$ is concave in p.



Joint entropy and Conditional entropy

Joint Entropy

- The **joint entropy of ensembles** X, Y is

$$H(X,Y) = \sum_{x \in \mathcal{A}_X} \sum_{y \in \mathcal{A}_Y} p(x,y) \log_2 \frac{1}{p(x,y)}.$$

- The joint entropy of two ensemble is the expected value of the information content of each of their joint outcomes:

$$H(X,Y) = \sum_{x \in \mathcal{A}_X} \sum_{y \in \mathcal{A}_Y} p(x,y) h(x,y).$$

- **Additivity in case of independence:** Entropy is additive for independent ensembles:

$$H(X,Y) = H(X) + H(Y) \quad \text{iff} \quad X \text{ and } Y \text{ are independent.}$$

Proof. We already had a proof for this in one of our homework assignments. In this lecture, however, we will provide an alternative proof soon. □

Conditional entropy

- We will now define a measure of the uncertainty about X given that the value y of a correlated ensemble Y is known.
- The **conditional entropy of an ensemble** X **given a result** $y \in \mathcal{A}_Y$ is

$$H(X \mid Y = y) = \sum_{x \in \mathcal{A}_X} p(x \mid Y = y) \log_2 \frac{1}{p(x \mid Y = y)}.$$

Conditional entropy

- We can also define a measure of the uncertainty about an ensemble X given that an ensemble Y is known.
- The **conditional entropy of an ensemble X given an ensemble Y** is

$$
\begin{aligned}
H(X \mid Y) &= \sum_{y \in \mathcal{A}_Y} p(y) H(X \mid Y = y) \\
&= \sum_{y \in \mathcal{A}_Y} p(y) \left(\sum_{x \in \mathcal{A}_X} p(x \mid y) \log_2 \frac{1}{p(x \mid y)} \right) \\
&= \sum_{x \in \mathcal{A}_X} \sum_{y \in \mathcal{A}_Y} p(x, y) \log_2 \frac{1}{p(x \mid y)}
\end{aligned}
$$

Conditional entropy

- **Theorem** If X and Y are independent, then $H(X \mid Y) = H(X)$.

Proof.

$$\begin{aligned}
H(X \mid Y) &= \sum_{x \in \mathcal{A}_X} \sum_{y \in \mathcal{A}_Y} p(x, y) \log_2 \frac{1}{p(x \mid y)} && \text{(by definition)} \\
&= \sum_{x \in \mathcal{A}_X} \sum_{y \in \mathcal{A}_Y} p(x) p(y) \log_2 \frac{1}{p(x)} && \text{(since } X \text{ and } Y \text{ are independent)} \\
&= \sum_{x \in \mathcal{A}_X} p(x) \log_2 \frac{1}{p(x)} \sum_{y \in \mathcal{A}_Y} p(y) && \text{(reorganizing the summations)} \\
&= \sum_{x \in \mathcal{A}_X} p(x) \log_2 \frac{1}{p(x)} \cdot 1 && (\textstyle\sum_{y \in \mathcal{A}_Y} p(y) = 1) \\
&= H(X) && \text{(by definition)}
\end{aligned}$$

□

Chain rules

- Chain rules tells us how to decompose a quantity that is a function of a joint distribution into a sum of functions on each individual component in the distribution.

- For instance, we are already familiar with the **chain rule of probabilities**:

  $$\begin{aligned} p(x, y) &= p(x)p(y \mid x) \\ &= p(y)p(x \mid y), \end{aligned}$$

  which intuitively means that the probability of results x and y happening together can be decomposed into the chain product of the probability of x happening and the probability of y happening given that x has happened.

- We will now derive chain rules for functions of dependent ensembles.

Chain rule for information content

- **Theorem (Chain rule of information content.)**

$$\begin{aligned} h(x, y) &= h(x) + h(y \mid x) \\ &= h(y) + h(x \mid y). \end{aligned}$$

Proof.

$$\begin{aligned} h(x, y) &= \log_2 \frac{1}{p(x, y)} && \text{(by def. of information content)} \\ &= \log_2 \frac{1}{p(x)p(y \mid x)} && \text{(chain rule for probabilities)} \\ &= \log_2 \frac{1}{p(x)} + \log_2 \frac{1}{p(y \mid x)} \\ &= h(x) + h(y \mid x) && \text{(by def. of information content)} \end{aligned}$$

The proof of $h(x, y) = h(y) + h(x \mid y)$ is analogous. □

Chain rule for entropy

- **Theorem** **(Chain rule for entropy.)**

$$\begin{aligned} H(X, Y) &= H(X) + H(Y \mid X) \\ &= H(Y) + H(X \mid Y) \end{aligned}$$

Proof.

$$\begin{aligned} H(X, Y) &= \sum_{x \in \mathcal{A}_X} \sum_{y \in \mathcal{A}_Y} p(x, y) h(x, y) && \text{(by definition)} \\ &= \sum_{x \in \mathcal{A}_X} \sum_{y \in \mathcal{A}_Y} p(x, y) \left[h(x) + h(y \mid x) \right] && \text{(chain rule for } h\text{)} \\ &= \sum_{x \in \mathcal{A}_X} \sum_{y \in \mathcal{A}_Y} \left[p(x, y) h(x) + p(x, y) h(y \mid x) \right] && \text{(distributivity)} \\ &= \sum_{x \in \mathcal{A}_X} \sum_{y \in \mathcal{A}_Y} p(x, y) h(x) + \sum_{x \in \mathcal{A}_X} \sum_{y \in \mathcal{A}_Y} p(x, y) h(y \mid x) && \text{(splitting the sums (*))} \end{aligned}$$

Chain rule for entropy

- **Proof.** (Continued)

Note that the first term in the sum in Equation (*) can be rewritten as

$$
\begin{aligned}
\sum_{x \in \mathcal{A}_X} \sum_{y \in \mathcal{A}_Y} p(x,y) h(x) &= \sum_{x \in \mathcal{A}_X} \sum_{y \in \mathcal{A}_Y} p(x) p(y \mid x) h(x) && \text{(chain rule of prob.)} \\
&= \sum_{x \in \mathcal{A}_X} p(x) h(x) \sum_{y \in \mathcal{A}_Y} p(y \mid x) && \text{(moving constants out of the summation)} \\
&= \sum_{x \in \mathcal{A}_X} p(x) h(x) \cdot 1 && \text{(since } \textstyle\sum_{y \in \mathcal{A}_Y} p(y \mid x) = 1\text{)} \\
&= H(X) && \text{(by def. of entropy (**)),}
\end{aligned}
$$

and the second term in the sum in Equation (*) can be rewritten as

$$
\sum_{x \in \mathcal{A}_X} \sum_{y \in \mathcal{A}_Y} p(x,y) h(y \mid x) = H(Y \mid X) \qquad \text{(by def. of conditional entropy (***)).}
$$

Chain rule for entropy

- **Proof.** (Continued)

  Now we can substitute Equations (**) and (***) in Equation (*) to obtain

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

  The proof of $H(X, Y) = H(Y) + H(X|Y)$ is analogous.

  □

Chain rule for entropy

- We can use what we saw so far in this lecture to prove the additivity of joint entropy in case of independence.

- **Corollary** Entropy is additive for independent ensembles:

$$H(X, Y) = H(X) + H(Y) \quad \text{iff} \quad X \text{ and } Y \text{ are independent.}$$

Proof. The result follows from the chain rule for entropy, which states that $H(X, Y) = H(X) + H(Y \mid X)$, and from the fact that if X and Y are independent, then $H(Y \mid X) = H(Y)$. □

Chain rule for conditional entropy

- The chain rule for entropy can be easily extended to conditional entropies.

- **Theorem (Chain rule for conditional entropy.)**

$$\begin{aligned} H(X, Y \mid Z) &= H(X \mid Z) + H(Y \mid X, Z) \\ &= H(Y \mid Z) + H(X \mid Y, Z) \end{aligned}$$

Proof. This is a corollary of the chain rule, just apply the same reasoning in the proof for the basic chain rule. □

General chain rule for entropy

- The chain rule for entropy can be extended to multiple ensembles as follows.

- **Theorem** The **general form of the chain rule for entropy** is

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i \mid X_1 \ldots, X_{i-1}).$$

Proof. (Sketch.) Let us consider the case of three ensembles X_1, X_2, X_3. To expand $H(X_1, X_2, X_3)$ we can consider X_2, X_3 as a single, joint ensemble Y, and apply the basic chain rule for entropy of two ensembles:

$$H(X_1, \underbrace{X_2, X_3}_{Y}) = H(X_1) + H(\underbrace{X_2, X_3}_{Y} \mid X_1).$$

General chain rule for entropy

- **Proof.** (Continued)

  Now we can use the chain rule for conditional entropy on $H(X_2, X_3 \mid X_1)$:

  $$H(X_2, X_3 \mid X_1) = H(X_2 \mid X_1) + H(X_3 \mid X_1, X_2).$$

  And then we substitute the above result in the previous equation, to obtain

  $$H(X_1, X_2, X_3) = H(X_1) + H(X_2 \mid X_1) + H(X_3 \mid X_1, X_2),$$

  which is the desired result for three ensembles.

  The general result for n ensembles can be shown by induction. □

Joint and conditional entropies - Example

- $\boxed{\text{Example 1}}$ **(Conditional entropies.)** Consider the joint ensemble X, Y distributed as follows.

$\mathcal{P}_{X,Y}$	y_1	y_2	y_3	y_4
x_1	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{4}$
x_2	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{16}$	0
x_3	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{16}$	0
x_4	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{16}$	0

The marginal distributions are

$$\mathcal{P}_X = \left\{ \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8} \right\},$$

and

$$\mathcal{P}_Y = \left\{ \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4} \right\}.$$

Joint and conditional entropies - Example

- $\boxed{\text{Example 1}}$ (Continued)

We can, then, calculate the entropies

$$
\begin{aligned}
H(X) &= -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{8}\log_2\frac{1}{8} - \frac{1}{8}\log_2\frac{1}{8} \\
&= \frac{7}{4} \text{ bits,}
\end{aligned}
$$

and

$$
\begin{aligned}
H(Y) &= -\frac{1}{4}\log_2\frac{1}{4} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{4}\log_2\frac{1}{4} \\
&= 2 \text{ bits,}
\end{aligned}
$$

and

$$
\begin{aligned}
H(X,Y) &= \sum_{x\in\mathcal{A}_X}\sum_{y\in\mathcal{A}_Y} p(x,y)\log_2\frac{1}{p(x,y)} \\
&= \frac{27}{8} \text{ bits.}
\end{aligned}
$$

Joint and conditional entropies - Example

- $\boxed{\text{Example 1}}$ (Continued)

To calculate the conditional entropy $H(X \mid Y)$ we first have to calculate the conditional distributions $p(X = x \mid Y = y)$ for each pair of values x, y.

The values of $p(x \mid y)$ are shown in the following table.

$\mathcal{P}_{X \mid Y}$	y_1	y_2	y_3	y_4
x_1	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	1
x_2	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	0
x_3	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{4}$	0
x_4	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{4}$	0

Note that <u>each column</u> in this table is a probability distribution on X given a fixed value $Y = y$.

Joint and conditional entropies - Example

- $\boxed{\text{Example 1}}$ (Continued)

  Now we can calculate

$$
\begin{aligned}
H(X \mid Y) &= \sum_{y \in \mathcal{A}_Y} p(y) H(X \mid Y = y) \\
&= p(y_1) H(X \mid Y = y_1) + p(y_2) H(X \mid Y = y_2) + \\
&\quad p(y_3) H(X \mid Y = y_3) + p(y_4) H(X \mid Y = y_4) \\
&= \frac{1}{4} H\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right) + \frac{1}{4} H\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{8}, \frac{1}{8}\right) + \\
&\quad \frac{1}{4} H\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) + \frac{1}{4} H(1, 0, 0, 0) \\
&= \frac{11}{8} \text{ bits.}
\end{aligned}
$$

Joint and conditional entropies - Example

- $\boxed{\text{Example 1}}$ (Continued)

Similarly, we can calculate the conditional entropy $H(Y \mid X)$ by first determining the conditional distributions $p(Y = y \mid X = x)$ for each pair of values x, y.

The values of $p(y \mid x)$ are shown in the following table.

$\mathcal{P}_{Y \mid X}$	y_1	y_2	y_3	y_4
x_1	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{2}$
x_2	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	0
x_3	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$	0
x_4	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$	0

Note that each row in this table is a probability distribution on Y given a fixed value $X = x$.

Joint and conditional entropies - Example

- $\boxed{\text{Example 1}}$ (Continued)

Now we can calculate

$$
\begin{aligned}
H(Y \mid X) &= \sum_{x \in \mathcal{A}_X} p(x) H(Y \mid X = x) \\
&= p(x_1) H(Y \mid X = x_1) + p(x_2) H(Y \mid X = x_2) + \\
&\quad p(x_3) H(Y \mid X = x_3) + p(x_4) H(Y \mid X = x_4) \\
&= \frac{1}{2} H\left(\frac{1}{4}, \frac{1}{8}, \frac{1}{8}, \frac{1}{2}\right) + \frac{1}{4} H\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}, 0\right) + \\
&\quad \frac{1}{4} H\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{2}, 0\right) + \frac{1}{8} H\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{2}, 0\right) \\
&= \frac{13}{8} \text{ bits.}
\end{aligned}
$$

Joint and conditional entropies - Example

- $\boxed{\text{Example 1}}$ (Continued)

  Note that this example shows the following very important property of conditional entropy.

  Conditional entropy is not symmetric in general:

$$H(X \mid Y) \neq H(Y \mid X).$$

◁

Conditional entropy - Challenges

- $\boxed{\text{Example 2}}$ Show that $H(Y \mid X) = 0$ if, and only if, Y is a function of X.

  Solution. Homework! $\triangleleft$

- $\boxed{\text{Example 3}}$ Let X be a random variable taking on a finite number of values.

  What is the most precise inequality relationship ($>$, $\geq$, $<$, $\leq$, $=$, or $\neq$) between $H(X)$ and $H(Y)$ if

  a) $Y = 2^X$?

  b) $Y = \cos X$?

  Solution. Homework! $\triangleleft$

Mutual information

Mutual information

- The mutual information is a measure of how much information two ensembles share.

  It can also work as a measure of how much information one ensemble carry about the other.

- The **mutual information between X and Y** is defined as

$$I(X;Y) = H(X) - H(X \mid Y).$$

- Intuitively, the amount of information Y carries about X is by how much the knowledge of Y reduces the uncertainty about X:

$$\overbrace{\left(\begin{array}{c}\text{information } Y\\ \text{carries about } X\end{array}\right)}^{I(X;Y)} = \overbrace{\left(\begin{array}{c}\text{prior uncertainty}\\ \text{about } X\end{array}\right)}^{H(X)} - \overbrace{\left(\begin{array}{c}\text{uncertainty about } X\\ \text{once } Y \text{ is konwn}\end{array}\right)}^{H(X \mid Y)}$$

Mutual information - Example

- $\boxed{\text{Example 4}}$ **(Mutual information.)** Consider again the joint ensemble X, Y distributed as follows.

$\mathcal{P}_{X,Y}$	y_1	y_2	y_3	y_4
x_1	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{4}$
x_2	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{16}$	0
x_3	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{16}$	0
x_4	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{16}$	0

Let us calculate:

- the information Y carries about X, given by $I(X;Y)$, and
- the information X carries about Y, given by $I(Y;X)$.

Mutual information - Example

- $\boxed{\text{Example 4}}$ (Continued)

  We calculated in a previous example that

$$H(X) = \frac{7}{4} \text{ bits} \qquad \text{and} \qquad H(X \mid Y) = \frac{11}{8}.$$

So we can calculate that

$$\begin{aligned} I(X;Y) &= H(X) - H(X \mid Y) \\ &= \frac{7}{4} - \frac{11}{8} \\ &= \frac{3}{8} \text{ bits,} \end{aligned}$$

and conclude that Y carries 3/8 bits of information about X.

Mutual information - Example

- $\boxed{\text{Example 4}}$ (Continued)

We also calculated in a previous example that

$$H(Y) = 2 \text{ bits} \qquad \text{and} \qquad H(Y \mid X) = \frac{13}{8} \text{ bits.}$$

Hence

$$\begin{aligned} I(Y; X) &= H(Y) - H(Y \mid X) \\ &= 2 - \frac{13}{8} \\ &= \frac{3}{8} \text{ bits,} \end{aligned}$$

and conclude that X carries $3/8$ bits of information about Y.

In this example, we have $I(X; Y) = I(Y; X)$. Is this a coincidence? ◁

Mutual information - Properties

- We will soon show the following important properties of mutual information:

  1. **Symmetry**: $I(X;Y) = I(Y;X)$.
  2. **Non-negativity**: $I(X;Y) \geq 0$.
  3. **Upper-bound**: $I(X;Y) \leq \min(H(X), H(Y))$.

- To prove such properties, though, we will first derive a relation among mutual information and relative entropy.

Relation between mutual information and relative entropy

Relative entropy

- We have seen that the entropy $H(X)$ is a measure of uncertainty of the ensemble X.

  The entropy $H(X)$ can also be seen as measure of the amount of information (in bits) necessary to describe an ensemble X.

- When we studied data compression, we saw that the Kullback-Liebler divergence $D_{KL}(p \parallel q)$ is a measure of “distance” between two distributions p and q.

  More precisely, we saw that $D_{KL}(p \parallel q)$ is a measure of the inefficiency of representing an ensemble with real probability distribution p by using a guessed probability distribution q.

- Here we will revisit the Kullback-Liebler divergence, under one of its other names: relative entropy.

Relative entropy

- The **relative entropy** (or **Kullback-Liebler divergence**) $D_{KL}(p \parallel q)$ of the probability mass function p with respect to the probability mass function q, both defined over a set $\mathcal{A}_X$, is defined as

  $$D_{KL}(p \parallel q) = \sum_{x \in \mathcal{A}_X} p(x) \log \frac{p(x)}{q(x)},$$

  with the convention that

  $$0 \cdot \log_2 \frac{0}{0} = 0, \qquad 0 \cdot \log_2 \frac{0}{q} = 0 \qquad \text{and} \qquad p \cdot \log_2 \frac{p}{0} = \infty.$$

- Note that if there is any value x s.t. $p(x) > 0$ but $q(x) = 0$, the KL-divergence is $D_{KL}(p \parallel q) = \infty$.

Relative entropy - Properties

- **Theorem (Non-negativity.)**

$$D_{KL}(p \parallel q) \geq 0,$$

with equality only if $p(x) = q(x)$ for all $x \in \mathcal{X}$.

Proof. The proof uses Jensen's inequality and the fact that $\log_2(x)$ is a concave function, and is left as an exercise for the student.

(The solution is given in Exercise 2.26 on MacKay's book.)

Relative entropy - Properties

- **Theorem (Non-symmetry.)** In general, $D_{KL}(p \parallel q) \neq D_{KL}(q \parallel p)$.

Proof. Consider $p = \left\{\frac{1}{2}, \frac{1}{2}\right\}$ and $q = \left\{\frac{3}{4}, \frac{1}{4}\right\}$. Then

$$\begin{aligned} D_{KL}(p \parallel q) &= \frac{1}{2} \log_2 \frac{1/2}{3/4} + \frac{1}{2} \log_2 \frac{1/2}{1/4} \\ &= 1 - \frac{1}{2} \log_2 3 \\ &= 0.2075 \text{ bits}, \end{aligned}$$

and

$$\begin{aligned} D_{KL}(q \parallel p) &= \frac{3}{4} \log_2 \frac{3/4}{1/2} + \frac{1}{4} \log_2 \frac{1/4}{1/2} \\ &= \frac{3}{4} \log_2 3 - 1 \\ &= 0.1887 \text{ bits}. \end{aligned}$$

□

Relation between mutual information and relative entropy

- The following result shows that the mutual information of two ensembles X and Y is the error you would commit by approximating their joint distribution $p(X, Y)$ by the product $p(X)p(Y)$ of their marginal distributions.

- **<u>Theorem</u> (Relation among mutual information and relative entropy.)**

$$\begin{aligned} I(X;Y) &= D_{KL}(p(x,y) \parallel p(x)p(y)) \\ &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}. \end{aligned}$$

(Note that if X and Y are independent, you commit no error by approximating their joint by the product of their marginals, and their mutual information is 0.)

Relation between mutual information and relative entropy

- **Proof.**

$$\begin{aligned} D_{KL}(p(x,y) \parallel p(x)p(y)) &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\ &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \left[\log p(x,y) - \log p(x) - \log p(y)\right] \\ &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log p(x,y) - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log p(x) - \\ & \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log p(y) \qquad \textbf{(*)} \end{aligned}$$

Note that the first term in the sum in Equation (*) can be rewritten as

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

Relation between mutual information and relative entropy

- **Proof.** (Continued)

The second term in the sum in Equation (*) can be rewritten as

$$\begin{aligned}
\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x) &= \sum_{x\in\mathcal{X}} \log p(x) \sum_{y\in\mathcal{Y}} p(x,y) && \text{(moving constants out of the summation)} \\
&= \sum_{x\in\mathcal{X}} (\log p(x))\, p(x) && \left(\textstyle\sum_{y\in\mathcal{Y}} p(x,y) = p(x)\right) \\
&= -H(X), && (***)
\end{aligned}$$

and the third term in the sum in Equation (*) can be rewritten as

$$\begin{aligned}
\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y) &= \sum_{y\in\mathcal{Y}} \log p(y) \sum_{x\in\mathcal{X}} p(x,y) && \text{(moving constants out of the summation)} \\
&= \sum_{y\in\mathcal{Y}} (\log p(y))\, p(y) && \left(\textstyle\sum_{x\in\mathcal{X}} p(x,y) = p(y)\right) \\
&= -H(Y). && (****)
\end{aligned}$$

Relation between mutual information and relative entropy

- **Proof.** (Continued)

And by replacing Equations (**), (***), and (****) in Equation (*) we get

$$D_{KL}(p(x,y) \parallel p(x)p(y)) = H(X) + H(Y) - H(X,Y) \qquad (*****)$$

To finish the proof we show that $H(X) + H(Y) - H(X,Y) = I(X;Y)$.

$$\begin{aligned} H(X) + H(Y) - H(X,Y) &= H(X) + H(Y) - (H(X) + H(Y \mid X)) && \text{(by the chain rule)} \\ &= H(Y) - H(Y \mid X) \\ &= I(X;Y) && \text{(by defnition),} \end{aligned}$$

And we can conclude that $D_{KL}(p(x,y) \parallel p(x)p(y)) = I(X;Y)$. □

Mutual information and Conditional entropy - Properties

- The relation between mutual information and relative entropy can be used to prove properties of mutual information and of conditional entropy.

- **Theorem (Symmetry of mutual information.)**

$$I(X;Y) = I(Y;X)$$

Proof.

$$\begin{aligned} I(X;Y) &= D_{KL}(p(x,y) \parallel p(x)p(y)) \\ &= D_{KL}(p(y,x) \parallel p(y)p(x)) \\ &= I(Y;X) \end{aligned}$$

□

Mutual information and Conditional entropy - Properties

- **Theorem (Non-negativity of mutual information.)**

$$I(X;Y) \geq 0,$$

with equality if, and only if, X and Y are independent.

Proof.

$$\begin{aligned} I(X;Y) &= D_{KL}(p(x,y) \parallel p(x)p(y)) \\ &\geq 0 && \text{(by non-neg. of rel. entropy).} \end{aligned}$$

Moreover, $D_{KL}(p(x,y) \parallel p(x)p(y)) = 0$ iff $p(x,y) = p(x)p(y)$, that is, iff X and Y are independent.

□

Mutual information and Conditional entropy - Properties

- **Theorem** **"Information can't hurt."** For any two ensembles X and Y, we have

$$H(X \mid Y) \leq H(X),$$

  with equality if, and only if, X and Y are independent.

  Proof. Because $I(X;Y) \geq 0$ we have that $H(X) - H(X \mid Y) \geq 0$, and, hence, $H(X) \geq H(X \mid Y)$.

  Moreover, by the previous result we know that $I(X;Y) = 0$ iff X and Y are independent, hence $H(X \mid Y) = H(X)$ iff X and Y are independent. □

- This property states that conditioning reduces entropy, or, in other words, that if you have some uncertainty about X, knowing Y can never increase your uncertainty about X.

Mutual information and Conditional entropy - Properties

- **Theorem (Upper-bound on mutual information.)**

$$I(X;Y) \leq \min\left(H(X), H(Y)\right).$$

Proof. Since $I(X;Y) = H(X) - H(X \mid Y)$ and $H(X) \geq H(X \mid Y)$, we must have $I(X;Y) \leq H(X)$.

Similarly, since $I(X;Y) = H(Y) - H(Y \mid X)$ and $H(Y) \geq H(Y \mid X)$, we must have $I(X;Y) \leq H(Y)$.

Putting together the facts that $I(X;Y) \leq H(X)$ and that $I(X;Y) \leq H(Y)$, we get that $I(X;Y) \leq \min\left(H(X), H(Y)\right)$. □

Conditional mutual information

- The conditional mutual information represents the amount of information X conveys about Y in a setting where Z is already part of the background knowledge.

  Formally, the **conditional mutual information between X and Y given Z**, given by

$$\underbrace{I(X;Y \mid Z)}_{\left(\begin{array}{c}\text{information } X \text{ and } Y\\ \text{share given}\\ Z \text{ is known}\end{array}\right)} = \underbrace{H(X \mid Z)}_{\left(\begin{array}{c}\text{uncertainty about}\\ X \text{ given}\\ Z \text{ is known}\end{array}\right)} - \underbrace{H(X \mid Y, Z)}_{\left(\begin{array}{c}\text{uncertainty about } X\\ \text{given } Z \text{ and } Y\\ \text{are known}\end{array}\right)}$$

Chain rule for mutual information

- The chain rule for mutual information tells us how to compose the information gained by a series of ensembles.

  Formally, the **chain rule for mutual information** is

$$\underbrace{I(X_1, X_2; Y)}_{\left(\begin{array}{c}\text{information } Y \text{ shares}\\ \text{with } X_1 \text{ and } X_2\end{array}\right)} = \underbrace{I(X_1; Y)}_{\left(\begin{array}{c}\text{information } Y\\ \text{shares with } X_1\end{array}\right)} + \underbrace{I(X_2; Y \mid X_1)}_{\left(\begin{array}{c}\text{information } Y \text{ shares}\\ \text{with } X_2 \text{ given } X_1 \text{ is known}\end{array}\right)}$$

- The **general form of the chain rule for mutual information** is

$$I(X_1, X_2, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y \mid X_1, \ldots, X_{i-1}).$$

Mutual information and Conditional entropy - Properties

- A set of random variables X_1, X_2, ..., X_n form a **Markov Chain**, denoted by

  $$X_1 \rightarrow X_2 \rightarrow \ldots \rightarrow X_n,$$

  when every X_i depends only on X_{i-1}.

  Formally, for all $1 \leq i \leq n$:

  $$p(x_n \mid x_1, x_2, \ldots, x_{n-1}) = p(x_n \mid x_{n-1}).$$

- It can be easily proven that when $X_1 \rightarrow X_2 \rightarrow \ldots \rightarrow X_n$ form a Markov chain, then:

  $$\begin{aligned} p(x_1, x_2, \ldots, x_n) = & \prod_{i=1}^{n} p(x_i \mid x_{i-1}) \\ = & \; p(x_1) \cdot p(x_2 \mid x_1) \cdot \ldots \cdot p(x_i \mid x_{i-1}) \cdot \ldots \cdot p(x_n \mid x_{n-1}). \end{aligned}$$

Mutual information and Conditional entropy - Properties

- **Theorem (Data-processing inequality.)** If $X \to Y \to Z$ forms a Markov chain, then

$$I(X;Y) \geq I(X;Z).$$

  Proof. This proof is part of your homework assignment for this lecture.

- The data-processing inequality can be used to show that no clever manipulation of the data can improve the inferences that can be made from the data.

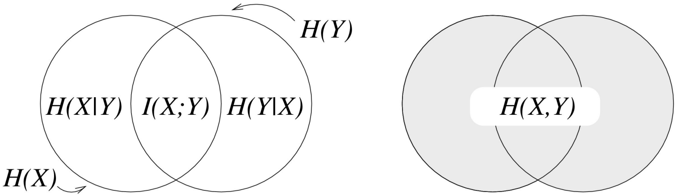
  In other words, the data-processing inequality states that:

  "Information can be destroyed during computation, but never created!"

Visual representation of relationship between entropies

Visual representation of relationship between entropies

- Some people represent the relationship between entropy, conditional entropy, joint entropy and mutual information as a Venn diagram:



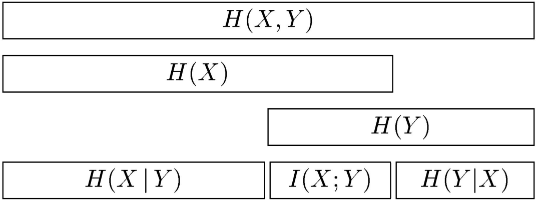
This diagram represents some essential relations among entropies:

1. $I(X;Y) = H(X) - H(X \mid Y)$,
2. $I(X;Y) = H(Y) - H(Y \mid X)$,
3. $I(X;Y) = H(X) + H(Y) - H(X,Y)$,
4. $\ldots$

Visual representation of relationship between entropies

- Even if the Venn diagram works well for two ensembles X and Y, it be misleading if we have three ensembles X, Y and Z.

- A preferable visual representation uses bars:



1. $I(X;Y) = H(X) - H(X \mid Y)$,
2. $I(X;Y) = H(Y) - H(Y \mid X)$,
3. $I(X;Y) = H(X) + H(Y) - H(X,Y)$,
4. $\ldots$

Challenges

- $\boxed{\text{Example 5}}$ Let X, Y, and Z form a joint ensemble.

  First explain in English what the following inequalities intuitively mean, and then prove them.

  a) $H(X, Y \mid Z) \geq H(X \mid Z)$.

  b) $I(X, Y; Z) \geq I(X; Z)$.

  c) $H(X, Y, Z) - H(X, Y) \leq H(X, Z) - H(X)$.

Solution. Homework!

◁

Summary

Entropy

- **Definition of entropy:** The entropy $H(X)$ of a discrete ensemble X is defined by

$$H(X) = -\sum_{x \in \mathcal{A}_X} p(x) \log p(x).$$

- **Properties of entropy:**

1. **Bounds for entropy:** $0 \leq H(X) \leq \log |\mathcal{A}_X|$.
2. **Changing the base of the entropy:** $H_b(X) = (log_a b) \cdot H_a(X)$.
3. **Conditioning reduces entropy:** For any two random variables X and Y, we have

$$H(X \mid Y) \leq H(X),$$

   with equality if, and only if, X and Y are independent.

4. **Concavity:** $H(p)$ is concave in p.

Mutual information

- **Definition of mutual information:** The mutual information between two ensembles X and Y is defined as

$$I(X;Y) = H(X) - H(X \mid Y).$$

- **Conditional mutual information:** The mutual information between two ensembles X and Y given that the ensemble Z is known is given by

$$I(X;Y \mid Z) = H(X \mid Z) - H(X \mid Y, Z).$$

- **Properties of mutual information:**

1. **Symmetry:** $I(X;Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X) = I(Y;X)$.
2. **Non-negativity:** $I(X;Y) \geq 0$, with equality if, and only if, X and Y are independent.
3. **Upper-bound:** $I(X;Y) \leq \min(H(X), H(Y))$.

Relative entropy

- **Definition of relative Entropy** (a.k.a., **Kullback-Liebler Divergence**): The relative entropy $D_{KL}(p \parallel q)$ of the probability mass function p with respect to the probability mass function q is defined by

$$D_{KL}(p \parallel q) = \sum_{x} p(x) \log \frac{p(x)}{q(x)}.$$

- **Properties of relative entropy:**

  1. **Non-negativity:** $D_{KL}(p \parallel q) \geq 0$, with equality only if $p(x) = q(x)$ for all $x \in \mathcal{X}$.

  2. **Non-symmetry:** in general, $D_{KL}(p \parallel q) \neq D_{KL}(q \parallel p)$.

- **Relation between mutual information and relative entropy.**

$$I(X;Y) = D_{KL}(p(x,y) \parallel p(x)p(y)) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}.$$

Chain rules

- **Chain rule for entropy:**

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i \mid X_1, \ldots, X_{i-1}).$$

- **Chain rule for mutual information:**

$$I(X_1, X_2, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y \mid X_1, \ldots, X_{i-1}).$$

Data-processing inequality

- **Data-processing inequality:** If $X \rightarrow Y \rightarrow Z$ forms a Markov chain, then

$$I(X;Y) \geq I(X;Z).$$

Alternative expressions

- **Alternative expressions in terms of expectations:**

  1. Entropy: $H(X) = E_p \log \frac{1}{p(X)}$.

  2. Joint entropy: $H(X, Y) = E_p \log \frac{1}{p(X, Y)}$.

  3. Conditional entropy: $H(X \mid Y) = E_p \log \frac{1}{p(X \mid Y)}$.

  4. Mutual information: $I(X; Y) = E_p \log \frac{p(X, Y)}{p(X)p(Y)}$.

  5. Relative entropy: $D_{KL}(p \parallel q) = E_p \log \frac{p(X)}{q(X)}$.