Information Theory and Statistics
Lecture 1: Entropy and information

Łukasz Dębowski
ldebowsk@ipipan.waw.pl

Ph. D. Programme 2013/2014
Entropy of a random variable on a probability space is the fundamental concept of information theory developed by Claude Shannon in 1948.
Probability as a random variable

**Definition**

Let $X$ and $Y$ be discrete variables and $A$ be an event on a probability space $(\Omega, \mathcal{F}, P)$. We define $P(X)$ as a discrete random variable such that

$$P(X)(\omega) = P(X = x) \iff X(\omega) = x.$$ 

Analogously we define $P(X|Y)$ and $P(X|A)$ as

$$P(X|Y)(\omega) = P(X = x|Y = y) \iff X(\omega) = x \text{ and } Y(\omega) = y,$$

$$P(X|A)(\omega) = P(X = x|A) \iff X(\omega) = x,$$

where the conditional probability is $P(B|A) = P(B \cap A)/P(A)$ for $P(A) > 0$. 

---

Project co-financed by the European Union within the framework of the European Social Fund
We write $P(Y) = P(X_1, X_2, \ldots, X_n)$ for $Y = (X_1, X_2, \ldots, X_n)$.

**Definition (independence)**

We say that random variables $X_1, X_2, \ldots, X_n$ are independent if

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i).$$

Analogously, we say that random variables $X_1, X_2, X_3, \ldots$ are independent if $X_1, X_2, \ldots, X_n$ are independent for any $n$.

**Example**

Let $\Omega = [0, 1]$ be the unit section and let $P$ be the Lebesgue measure. Define real random variable $Y(\omega) = \omega$. If we consider its binary expansion $Y = \sum_{i=1}^{\infty} 2^{-i}Z_i$, where $Z_i : \Omega \to \{0, 1\}$, then

$$P(Z_1, Z_2, \ldots, Z_n) = 2^{-n} = \prod_{i=1}^{n} P(Z_i).$$
Definition (expectation)

We define the expectation of a real random variable $X$ as

$$
E_X := \int XdP.
$$

For discrete random variables we obtain

$$
E_X = \sum_{x: P(X = x) > 0} P(X = x) \cdot x.
$$
One of fundamental properties of the expectation is its additivity.

**Theorem**

Let $X, Y \geq 0$. We have

$$
\mathbb{E}(X + Y) = \mathbb{E}X + \mathbb{E}Y.
$$

**Remark:** The restriction $X, Y \geq 0$ is made because, e.g., for $\mathbb{E}X = \infty$ and $\mathbb{E}Y = -\infty$ the sum is undefined.
Some interpretation of entropy $H(X)$ is the average uncertainty carried by a random variable $X$. We expect that uncertainty adds for probabilistically independent sources. Formally, for $P(X, Y) = P(X)P(Y)$, we postulate $H(X, Y) = H(X) + H(Y)$. Because $\log(xy) = \log x + \log y$, the following definition comes as a very natural idea.

**Definition (entropy)**

The *entropy* of a discrete variable $X$ is defined as

$$H(X) := \mathbb{E} [− \log P(X)].$$

Traditionally, it is assumed that $\log$ is the logarithm to the base 2.

Because $\log P(X) \leq 0$, we put the minus sign in the definition (1) so that entropy be positive.
Equivalently, we have

\[ H(X) = - \sum_{x: P(X=x) > 0} P(X=x) \log P(X=x), \]

We can verify that for \( P(X, Y) = P(X)P(Y) \),

\[ H(X, Y) = \mathbb{E} [-\log P(X, Y)] = \mathbb{E} [-\log P(X) - \log P(X)] \]
\[ = \mathbb{E} [-\log P(X)] + \mathbb{E} [-\log P(X)] = H(X) + H(Y). \]
Entropy for a binary variable

Figure: Entropy $H(X) = -p \log p - (1 - p) \log(1 - p)$ for $P(X = 0) = p$ and $P(X = 1) = 1 - p$. 
The range of entropy in general

Because function \( f(p) = -p \log p \) is strictly positive for \( p \in (0, 1) \) and equals 0 for \( p = 1 \), it can be easily seen that:

**Theorem**

\[
H(X) \geq 0, \text{ whereas } H(X) = 0 \text{ if and only if } X \text{ assumes only a single value.}
\]

This fact agrees with the idea that constants carry no uncertainty. On the other hand, assume that \( X \) takes values \( x \in \{1, 2, \ldots, n\} \) with equal probabilities \( P(X = x) = 1/n \). Then

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

As we will see, \( \log n \) is the maximal value of \( H(X) \) if \( X \) assumes values in \( \{1, 2, \ldots, n\} \). That fact agrees with the intuition that the highest uncertainty occurs for uniformly distributed variables.

Project co-financed by the European Union within the framework of the European Social Fund
A discrete probability distribution is a function \( p : \mathbb{X} \rightarrow [0, 1] \) on a countable set \( \mathbb{X} \) such that \( p(x) \geq 0 \) and \( \sum_x p(x) = 1 \).

**Definition (entropy revisited)**

The entropy of a discrete probability distribution is denoted as

\[
H(p) := - \sum_{x: p(x) > 0} p(x) \log p(x).
\]

**Definition (KL divergence)**

*Kullback-Leibler divergence*, or *relative entropy* of probability distributions \( p \) and \( q \) is defined as

\[
D(p \parallel q) := \sum_{x: p(x) > 0} p(x) \log \frac{p(x)}{q(x)}.
\]
Convex and concave functions

Definition (convex and concave functions)

A real function $f : \mathbb{R} \to \mathbb{R}$ is convex if

$$p_1 f(x_1) + p_2 f(x_2) \geq f(p_1 x_1 + p_2 x_2)$$

for $p_i \geq 0$ and $p_1 + p_2 = 1$. Moreover, $f$ is called strictly convex if

$$p_1 f(x_1) + p_2 f(x_2) > f(p_1 x_1 + p_2 x_2)$$

for $p_i > 0$ and $p_1 + p_2 = 1$. We say that function $f$ is concave if $-f$ is convex, whereas $f$ is strictly concave if $-f$ is strictly convex.

Example

If function $f$ has a positive second derivative then it is strictly convex. Hence functions $h(x) = -\log x$ and $g(x) = x^2$ are strictly convex.
Theorem (Jensen inequality)

*If* \( f \) *is a convex function and* \( p \) *is a discrete probability distribution over real values then*

\[
\sum_{x: p(x) > 0} p(x)f(x) \geq f \left( \sum_{x: p(x) > 0} p(x) \cdot x \right).
\]

*Moreover, if* \( f \) *is strictly convex then*

\[
\sum_{x: p(x) > 0} p(x)f(x) = f \left( \sum_{x: p(x) > 0} p(x) \cdot x \right)
\]

*holds if and only if distribution* \( p \) *is concentrated on a single value.*

The proof proceeds by induction on the number of values of \( p \).
KL divergence is nonegative

**Theorem**

We have

\[ D(p||q) \geq 0, \]

where the equality holds if and only if \( p = q \).

**Proof**

By the Jensen inequality for \( f(y) = -\log y \), we have

\[
D(p||q) = - \sum_{x} p(x) \log \frac{q(x)}{p(x)} \geq - \log \left( \sum_{x} p(x) \frac{q(x)}{p(x)} \right) = - \log \left( \sum_{x} q(x) \right) \geq - \log 1 = 0.
\]
The maximum of entropy

**Theorem**

Let $X$ assume values in $\{1, 2, ..., n\}$. We have $H(X) \leq \log n$, whereas $H(X) = \log n$ if and only if $P(X = x) = 1/n$.

**Remark:** If the range of variable $X$ is infinite then entropy $H(X)$ may be infinite.

**Proof**

Let $p(x) = P(X = x)$ and $q(x) = 1/n$. Then

$$0 \leq D(p||q) = \sum_{x:p(x)>0} p(x) \log \frac{p(x)}{1/n} = \log n - H(X),$$

where the equality occurs if and only if $p = q$. 
The next important question is what is the behavior of entropy under conditioning. The intuition is that given additional information, the uncertainty should decrease. So should entropy. There are two distinct ways of defining conditional entropy.

**Definition (conditional entropy)**

*Conditional entropy* of a discrete variable $X$ given event $A$ is

$$H(X|A) := H(p) \text{ for } p(x) = P(X = x|A).$$

*Conditional entropy* of $X$ given a discrete variable $Y$ is defined as

$$H(X|Y) := \sum_{y: P(Y=y) > 0} P(Y = y)H(X|Y = y).$$

Both $H(X|A)$ and $H(X|Y)$ are nonnegative.
Minimum of conditional entropy

**Theorem**

\[ H(X|Y) = 0 \text{ holds if and only if } X = f(Y) \text{ for a certain function } f \text{ except for a set of probability 0.} \]

**Proof**

Observe that \( H(X|Y) = 0 \) if and only if \( H(X|Y = y) = 0 \) for all \( y \) such that \( P(Y = y) > 0 \). This holds if and only if given \( (Y = y) \) with \( P(Y = y) > 0 \), variable \( X \) is concentrated on a single value. Denoting this value as \( f(y) \), we obtain \( X = f(Y) \), except for the union of those sets \( (Y = y) \) which have probability 0.
We have

$$H(X|Y) = \mathbb{E} \left[ -\log P(X|Y) \right].$$

**Proof**

$$H(X|Y) = \sum_y P(Y = y) H(X|Y = y)$$

$$= - \sum_{x,y} P(Y = y) P(X = x|Y = y) \log P(X = x|Y = y)$$

$$= - \sum_{x,y} P(X = x, Y = y) \log P(X = x|Y = y)$$

$$= \mathbb{E} \left[ -\log P(X|Y) \right].$$
Because $P(Y)P(X|Y) = P(X, Y)$, by the previous result,

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

Hence

$$H(X, Y) \geq H(Y).$$
Inequality $H(X|A) \leq H(X)$ need not hold

**Example**

Let $P(X = 0|A) = P(X = 1|A) = 1/2$, whereas $P(X = 0|A^c) = 1$ and $P(X = 1|A^c) = 0$. Assuming $P(A) = 1/2$, we have $P(X = 0) = (1/2) \cdot (1/2) + (1/2) = 3/4$ and $P(X = 0) = (1/2) \cdot (1/2) = 1/4$ so

$$H(X) = -\frac{3}{4} \log \frac{3}{4} - \frac{1}{4} \log \frac{1}{4} = \log 4 - \frac{3}{4} \log 3 = 0.811....$$

On the other hand, we have $H(X|A) = \log 2 = 1$.

Despite that fact, $H(X|Y) \leq H(X)$ holds in general. Thus entropy decreases given additional information on average.
To show that $H(X)$ is greater than $H(X|Y)$, it is convenient to introduce another important concept.

**Definition (mutual information)**

*Mutual information* between discrete variables $X$ and $Y$ is

$$I(X; Y) := \mathbb{E} \left[ \log \frac{P(X, Y)}{P(X)P(Y)} \right].$$

We have $I(X; X) = H(X)$. Thus, entropy is sometimes called *self-information*. 
Mutual information is nonnegative

**Theorem**

We have

\[ I(X; Y) \geq 0, \]

where the equality holds if and only if \( X \) and \( Y \) are independent.

**Proof**

Let \( p(x, y) = P(X = x, Y = y) \) and \( q(x, y) = P(X = x)P(Y = y) \). Then we have

\[
I(X; Y) = \sum_{(x,y):p(x,y)>0} p(x, y) \log \frac{p(x, y)}{q(x, y)} = D(p||q) \geq 0
\]

with the equality exactly for \( p = q \).
Mutual information and entropy

By the definition of mutual information,

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

Hence

\[ H(X) + H(Y) \geq H(X, Y), \]
\[ H(X) \geq H(X|Y), \quad I(X; Y). \]

Moreover, we have \( H(X|Y) = H(Y) \) if \( X \) and \( Y \) are independent, which also agrees with intuition.
In a similar fashion to conditional entropy, we define:

**Definition (conditional mutual information)**

*Conditional mutual information* between discrete variables $X$ and $Y$ given event $A$ is

$$I(X; Y|A) := D(p||q) \text{ for } p(x, y) = P(X = x, Y = y|A)$$

and

$$q(x, y) = P(X = x|A)P(Y = y|A).$$

*Conditional mutual information* between discrete variables $X$ and $Y$ given variable $Z$ is defined as

$$I(X; Y|Z) := \sum_{z:P(Z=z)>0} P(Z = z)I(X; Y|Z = z).$$

Both $I(X; Y|A)$ and $I(X; Y|Z)$ are nonnegative.
Another formula for CMI

As in the case of conditional entropy, this proposition is true:

**Theorem**

We have

\[ I(X; Y | Z) := E \log \frac{P(X, Y | Z)}{P(X | Z)P(Y | Z)} \cdot \]
Conditional independence

**Definition (conditional independence)**

Variables $X_1, X_2, \ldots, X_n$ are *conditionally independent* given $Z$ if

$$P(X_1, X_2, \ldots, X_n | Z) = \prod_{i=1}^{n} P(X_i | Z).$$

Variables $X_1, X_2, X_3, \ldots$ are conditionally independent given $Z$ if $X_1, X_2, \ldots, X_n$ are conditionally independent given $Z$ for any $n$.

**Theorem**

*We have*

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

*with equality iff $X$ and $Y$ are conditionally independent given $Z$.***
Conditional independence (examples)

Example

Let $Y = f(Z)$ be a function of variable $Z$, whereas $X$ be an arbitrary variable. Variables $X$ and $Y$ are conditionally independent given $Z$. Indeed,

$$P(X = x, Y = y|Z = z) = P(X = x|Z = z) \cdot 1_{\{y = f(z)\}}$$

$$= P(X = x|Z = z)P(Y = y|Z = z).$$

Example

Let variables $X$, $Y$, and $Z$ be independent. Variables $U = X + Z$ and $W = Y + Z$ are conditionally independent given $Z$. Indeed,

$$P(U = u, W = w|Z = z) = P(X = u - z, Y = w - z)$$

$$= P(X = u - z)P(Y = w - z)$$

$$= P(U = u|Z = z)P(W = w|Z = z).$$
Definition (Markov chain)

A stochastic process \((X_i)_{i=-\infty}^{\infty}\) is called a Markov chain if

\[
P(X_i|X_{i-1}, X_{i-2}, \ldots, X_{i-n}) = P(X_i|X_{i-1})
\]

holds for any \(n\).

Example

For a Markov chain \((X_i)_{i=-\infty}^{\infty}\), variables \(X_i\) and \(X_k\) are conditionally independent given \(X_j\) if \(i \leq j \leq k\). Indeed, after some algebra we obtain

\[
P(X_k|X_i, X_j) = P(X_k|X_j), \quad \text{and hence}
\]

\[
P(X_i, X_k|X_j) = P(X_i|X_j)P(X_k|X_i, X_j) = P(X_i|X_j)P(X_k|X_j).
\]
Theorem

We have

\[ I(X; Y|Z) + I(X; Z) = I(X; Y, Z). \]

Remark: Hence, variables \( X \) and \((Y, Z)\) are independent iff \( X \) and \( Z \) are independent and \( X \) and \( Y \) are independent given \( Z \).

Proof

\[
I(X; Y|Z) + I(X; Z) = \mathbb{E} \left[ \log \frac{P(X, Y, Z)P(Z)}{P(X, Z)P(Y, Z)} \right] + \mathbb{E} \left[ \log \frac{P(X, Z)}{P(X)P(Z)} \right]
\]

\[
= \mathbb{E} \left[ \log \frac{P(X, Y, Z)}{P(X)P(Y, Z)} \right] = I(X; Y, Z).
\]
CMI and entropy

Theorem

If entropies $H(X)$, $H(Y)$, and $H(Z)$ are finite, we have

\[
H(X|Y) = H(X, Y) - H(Y),
\]
\[
I(X; Y) = H(X) - H(X|Y) = H(X) + H(Y) - H(X, Y),
\]
\[
I(X; Y|Z) = H(X|Z) + H(Y|Z) - H(X, Y|Z)
\]
\[
= H(X, Z) + H(Y, Z) - H(X, Y, Z) - H(Z),
\]

where all terms are finite and nonnegative.

The proof is left as an easy exercise.