**Chebyshev’s Inequality**

**Theorem:** Let $X$ be a random variable. For all $\varepsilon > 0$,

$$p(|X - E(X)| \geq \varepsilon) \leq \frac{1}{\varepsilon^2} \sigma^2(X).$$

**Proof:** Changing $X$ by a constant changes $E(X)$ by the same constant, doesn’t change $X - E(X)$, doesn’t change $\sigma^2(X)$.

Put $Y = X - E(X)$. Then $E(Y) = 0$. So our task is to prove that

$$p(|Y| \geq \varepsilon) \leq \frac{1}{\varepsilon^2} \sigma^2(Y)$$

given $E(Y) = 0$.

$$\sigma^2(Y) = E(Y^2) = \sum_y p(y) y^2 \geq \sum_{|y| \geq \varepsilon} p(y) y^2$$

$$\geq \sum_{|y| \geq \varepsilon} p(y) \varepsilon^2 = \varepsilon^2 p(|Y| \geq \varepsilon).$$

---

**Axioms for information**

The following rules are intuitively reasonable.

1. The information content of an event depends on its probability.
2. If $p(X = x)$ is 1 then $\text{Info}(X = x)$ is 0, and $\text{Info}(X = x) \to \infty$ as $p(X = x) \to 0^+$.
3. If $X$ and $Y$ are independent then

   \[ \text{Info}(X = x \text{ and } Y = y) \]

   should be equal to

   \[ \text{Info}(X = x) + \text{Info}(Y = y). \]

4. If $p(X = x)$ is $\frac{1}{2}$ then $\text{Info}(X = x)$ is defined to be “1 bit”.

---

**Reformulating these axioms**

Axiom 1 says that $\text{Info}(X = x) = S(p(X = x))$ for some nonnegative-valued function $S$ defined on the interval $(0, 1]$.

Axiom 2 says that

(2) \hspace{1cm} S(1) = 0 \text{ and } S(p) \to \infty \text{ as } p \to 0^+.

If $X$, $Y$ are independent then $p(X = x \text{ and } Y = y)$ equals the product $p(X = x)p(Y = y)$. So Axiom 3 says that

(3) \hspace{1cm} S(pq) = S(p) + S(q).

And Axiom 4 says that

(4) \hspace{1cm} S\left(\frac{1}{2}\right) = 1.
Determining the function

**Theorem:** The only function \( S \) satisfying (2), (3) and (4) above is \( S(p) = -\log_2(p) \).

**Proof:** Suppose that \( S \) satisfies (2), (3) and (4).

If \( t \in [0, \infty) \) then \( 2^t \geq 1 \) and so \( 0 < 2^{-t} \leq 1 \).

Define \( f : [0, \infty) \rightarrow \mathbb{R} \) by \( f(t) = S(2^{-t}) \).

Then
\[
f(t_1 + t_2) = S(2^{-t_1-t_2}) = S(2^{-t_1}2^{-t_2}) = S(2^{-t_1}) + S(2^{-t_2}) = f(t_1) + f(t_2).
\]

By an easy induction, \( f(nt) = nf(t) \) for all \( t \geq 0 \) and all \( n \in \mathbb{Z}^+ \).

And since \( f(1) = S(2^{-1}) = 1 \) we get \( f(n) = n \) for all \( n \in \mathbb{Z}^+ \).

So if \( t = \frac{n}{m} \) then \( mf(t) = f(mt) = f(n) = n \).

Hence \( f(\frac{n}{m}) = \frac{n}{m} \). That is, \( f(q) = q \) for all rational \( q \geq 0 \).

Proof continued

We defined \( f(t) = S(2^{-t}) \). Equivalently, \( S(p) = f(-\log_2 p) \).

We have shown that \( f(q) = q \) for all rational \( q \geq 0 \).

If \( t \leq t' \) then \( f(t) \leq f(t) + f(t' - t) = f(t') \). So \( f \) is increasing.

Let \( t \in \mathbb{R}^+ \). For any \( \varepsilon > 0 \) there is an \( m \in \mathbb{Z}^+ \) with \( \frac{1}{m} < \varepsilon \).

And \( \frac{n}{m} \leq t < \frac{n+1}{m} \) for some \( n \in \mathbb{Z} \).

So the rational numbers \( q_1 = \frac{n}{m} \) and \( q_2 = \frac{n+1}{m} \) differ by less than \( \varepsilon \) and satisfy \( q_1 \leq t < q_2 \).

Now
\[
q_1 = f(q_1) \leq f(t) < f(q_2) = q_2,
\]

and so \( t \) and \( f(t) \) differ by less than \( \varepsilon \).

But \( \varepsilon \) is an arbitrary positive number. So \( f(t) = t \).

So \( S(p) = f(\log_2 p) = -\log_2 p \) for all \( p \in (0, 1] \).

Entropy

If \( X \) is a random variable then \( -\log_2(p(X)) \) is a random variable giving the information content of each event \( x \).

The expectation of \( -\log_2(p(X)) \) is the amount of information you expect to get, on average, from the outcome of the experiment.

**Definition:** The entropy of a random variable \( X \), denoted by \( H(X) \), is its expected information content.

That is, \( H(X) = E(-\log_2(p(X))) \), the amount of information you expect to acquire by learning the outcome of \( X \).

**Note:** The greater your uncertainty about the outcome of an experiment, the more information you expect to acquire from it.

So entropy measures uncertainty.

Logarithmic growth

A result from 1st year calculus: \( \frac{\ln t}{t} \rightarrow 0 \) as \( t \rightarrow \infty \).

**Proof:** Note that \( 0 < 1/u \leq 1 \) for \( u \geq 1 \). So for \( t \geq 1 \),
\[
0 \leq \int_1^t \frac{du}{u} \leq \int_1^t du = t - 1 < t,
\]

and since \( \int_1^t \frac{du}{u} = \ln t \) we deduce that \( 0 \leq \frac{\ln t}{t} < 1 \).

Hence \( 0 \leq \frac{\ln t}{t} \leq \frac{2\ln \sqrt{t}}{\sqrt{t}\sqrt{t}} = \frac{2}{\sqrt{t}} \times \frac{\ln \sqrt{t}}{\sqrt{t}} \leq \frac{2}{\sqrt{t}} \rightarrow 0 \) as \( t \rightarrow \infty \).

Applying the Squeeze Law gives the result.

Putting \( p = 1/t \), this gives \( \lim_{p \to 0^+} p(-\ln p) = 0 \).

Dividing through by \( -\ln 2 \) this gives \( \lim_{p \to 0^+} p \log_2 p = 0 \).
**Jointly distributed variables**

Suppose that $X$ and $Y$ are jointly distributed (i.e. not necessarily independent) random variables.

Definition: The **joint entropy** of $X$ and $Y$ is the quantity

\[ H(X, Y) = E(-\log_2(p(X, Y))). \]

That is, the joint entropy is just the entropy of the random variable $(X, Y)$.

For example, let $X = \{r, s\}$ and $Y = \{a, b, c\}$ have the following joint distribution:

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>r</td>
<td>0.1</td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td>s</td>
<td>0.2</td>
<td>0.1</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Then $H(X, Y) = -0.1 \log(0.1) - 0.2 \log(0.2) - 0.1 \log(0.1) - 0.2 \log(0.2) - 0.1 \log(0.1) - 0.3 \log(0.3)$.

**Joint entropy for independent variables**

**Theorem:** If $X$ and $Y$ are independent then

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

**Proof:** Independence gives $p(x, y) = p(x)p(y)$, and so

\[ H(X, Y) = -\sum_x \sum_y p(x, y) \log(p(x, y)) \]

\[ = -\sum_x \sum_y p(x)p(y) \log(p(x)p(y)) \]

\[ = -\sum_x \sum_y p(x)p(y)(\log p(x) + \log p(y)) \]

\[ = -\sum_x \sum_y p(x)p(y) \log p(x) - \sum_x \sum_y p(x)p(y) \log p(y). \]

But $-\sum_x \sum_y p(x)p(y) \log p(x) = -\sum_x p(x) \log p(x) \sum_y p(y)$ which is $-\sum_x p(x) \log p(x) = H(X)$ (since $\sum_y p(y) = 1$).

Similarly $-\sum_p \sum_y p(x)p(y) \log p(y) = H(Y)$. 

**Two graphs**

Here is the graph of $-p \log_2 p$ against $p$ for $p \in (0, 1]$, and the graph of $\phi(p) = -p \log_2 p - (1 - p) \log_2(1 - p)$ against $p$, also for $p \in (0, 1]$.

The function $\phi$ gives the value of $H(X)$ for the experiment of tossing a coin that comes down heads with probability $p$.

There is less entropy – less uncertainty – when $p$ is close to 0 or 1 than when it is close to 0.5.
Suppose that $X$ and $Y$ are jointly distributed. For each fixed value $y$ of $Y$ the numbers $p(x|y) = p(x, y)/p(y)$ form a probability distribution for $X$. The entropy of this distribution is

$$f(y) = - \sum_x p(x|y) \log(p(x|y)).$$

We call this the \textit{conditional entropy of $X$ given $Y = y$}. This is a function of $Y$. Its expectation is called the \textit{conditional entropy of $X$ given $Y$} and is denoted by $H(X|Y)$. So

$$H(X|Y) = \sum_y p(y) f(y) = - \sum_{x,y} p(y)p(x|y) \log(p(x|y)).$$

We defined $H(X|Y) = - \sum_{x,y} p(y)p(x|y) \log(p(x|y))$.

The definition of $p(x|y)$ gives $p(y)p(x|y) = p(x, y)$. So

$$H(X|Y) = - \sum_{x,y} p(x, y) \log(p(x, y)/p(y))
= - \sum_{x,y} p(x, y)(\log p(x, y) - \log p(y))
= - \sum_{x,y} p(x, y) \log p(x, y) + \sum_y \sum_x p(x, y) \log p(y).$$

The first term here is $H(X, Y)$, and the second term is

$$\sum_y \sum_x p(x, y) \log p(y) = \sum_y \log p(y) \sum_x p(x, y).$$

Since $\sum_x p(x, y) = p(y)$ this is $\sum_y p(y) \log p(y) = -H(Y)$. So

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