ENTROPY

HARM DERKSEN

1. Probability

Let $X$ be a random variable. Suppose that $X$ can take a finite number of values, say \( \{x_1, x_2, \ldots, x_n\} \). The probability that $X$ is equal to $x_i$ is denoted by $P(X = x_i)$. For convenience, let us write $p_i = P(X = x_i)$. Then we must have

$$p_1 + p_2 + \cdots + p_n = 1$$

and $0 \leq p_i \leq 1$ for all $i$.

If the $x_i$ are real or complex numbers, then we can also define the expected value of $X$ by

$$E(X) = p_1x_1 + p_2x_2 + \cdots + p_nx_n.$$ 

Suppose that $X$ is a random variable that can take values \( \{x_1, \ldots, x_n\} \) and $Y$ is a random variable that can take values \( \{y_1, y_2, \ldots, y_m\} \) then $X$ and $Y$ are called independent if

$$P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j)$$

for all $i$ and $j$.

We also will need conditional probability. We define

$$P(X = x_i \mid Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)}$$

This can be thought of as the chance that $X$ has value $x_i$, under the assumption that $Y = y_j$. For more details, see any book on probability theory.

2. Entropy

Suppose again that $X$ is a random variable with values \( \{x_1, x_2, \ldots, x_n\} \) and probabilities \( \{p_1, p_2, \ldots, p_n\} \). Let $b > 1$ be a real number. We define the $b$-ary entropy of $X$ by

$$H_b(X) = H_b(p_1, p_2, \ldots, p_n) = -p_1 \log_b p_1 - p_2 \log_b p_2 - \cdots - p_n \log_b p_n.$$ 

The function $x \log_b(x)$ is a priori only defined for $x > 0$. Since $\lim_{x \downarrow 0} x \log_b(x) = 0$, we can extend $x \log_b(x)$ to a continuous function for all $x \geq 0$. Now (1) also makes sense if $p_i = 0$ for some $i$.

We call $H_2(X)$ the binary entropy, or just entropy, and we will just write $H(X)$ instead of $H_2(X)$ and log instead of $\log_2$. 

\[\text{Page 1}\]
We try to make a point here that $H_0(X)$ measures the “amount of uncertainty” or “the amount of information” of the random variable $X$. Let us first compute a few examples.

**Example 1.** Let $X$ be a random variable with values \{0, 1\}, and with $P(X = 0) = P(X = 1) = \frac{1}{2}$. For example, this could be the flipping of one fair coin. Then the uncertainty of $X$ is

$$H(X) = H\left(\frac{1}{2}, \frac{1}{2}\right) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = -\log_2\frac{1}{2} = 1.$$  
Suppose that $Y_n$ is the random variable of flipping $n$ coins. The values of $Y_n$ are all binary words of length $n$. These are $2^n$ possibilities. Each word has probability $\frac{1}{2^n}$. The uncertainty of $Y_n$ is

$$H(Y_n) = H(\frac{1}{2^n}, \ldots, \frac{1}{2^n}) = 2^n\left(-\frac{1}{2^n}\log_2\frac{1}{2^n}\right) = -\log_2\frac{1}{2^n} = n.$$  

**Example 2.** The chance that an egg will break if it falls from 4 feet above the ground is 99%. The uncertainty of whether the egg will break or not is

$$H\left(\frac{99}{100}, \frac{1}{100}\right) = -\frac{99}{100}\log_2\left(\frac{99}{100}\right) - \frac{1}{100}\log_2\left(\frac{1}{100}\right) \approx 0.0808 < 1.$$  
Indeed, the outcome of the egg experiment is much less uncertain than the flipping of a coin!

**Example 3.** In the land of Oz, the alphabet has only three letters, namely $a$, $b$ and $c$. In a given text, $a$ appears with probability $\frac{1}{2}$, and $b$ and $c$ each appear with probability $\frac{1}{4}$.

$$H(X) = H\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right) = \frac{1}{2}\log_2 2 + \frac{1}{4}\log_2 4 + \frac{1}{4}\log_2 4 = \frac{1}{2} + \frac{1}{2} + \frac{1}{2} = \frac{3}{2}.$$  
Suppose that $a$ is encoded by 0, $b$ is encoded by 10 and $c$ is encoded by 11. For example the text

$$abcaabac$$
translates to

$$010110010011$$

Given the encoded message, one can always find back the original Oz-message (think about this for a while why this is true). The expected length of one encoded letter is

$$\frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{4} \cdot 2 = \frac{3}{2}$$

which is in this case exactly equal to the entropy of a single letter. For a random text of length $n$, the encoded message will have expected length $\frac{3}{2}n$.

This example gives another interpretation of the entropy. The entropy $H(X)$ is roughly (but not exactly) the expected number of bits you need to encode the outcome of a random variable $X$. We probably will not give proofs or precise statements in this direction, but it may be useful to keep this in mind when you think of entropy.
Example 4. Suppose we have a computer text file which is a binary word of $n$ bits. Because it is a text file, certain patterns in the file are more likely than other patterns. Due to this, the uncertainty of a random text file of $n$ bits is smaller than $n$, say $a < n$. By clever encoding, the expected length of an encoded message may be close to $a$. This shows that the notion of entropy plays an important role in data compression (you may know commands like zip, compress or gzip on a computer). (Again, we will not prove these assertions here.)

3. Properties of Entropy

We list some properties of Entropy, together with an interpretation of the given formulas.

1. $H_b(p_1, \ldots, p_n)$ is continuous function for all $p_1, p_2, \ldots, p_n \in \mathbb{R}$ with $p_i \geq 0$ for all $i$ and $p_1 + p_2 + \cdots + p_n = 1$.

2. 
   $$H_b\left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right) < H_b\left(\frac{1}{n+1}, \frac{1}{n+1}, \ldots, \frac{1}{n+1}\right)$$

   This indeed makes sense. Suppose that $X$ is a random variable with values $\{x_1, x_2, \ldots, x_n\}$ and each value $x_i$ is equally likely with probability $\frac{1}{n}$, and $Y$ is a random variable with values $\{y_1, \ldots, y_{n+1}\}$ and each value $y_i$ is equally likely with probability $\frac{1}{n+1}$. Then $Y$ should have more “uncertainty” than $X$.

3. If $k_1, k_2, \ldots, k_m$ are positive integers with $k_1 + \cdots + k_m = n$, then
   
   $$H_b\left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right) = H_b\left(\frac{k_1}{n}, \frac{k_2}{n}, \ldots, \frac{k_m}{n}\right) + \sum_{i=1}^{m} \frac{k_i}{n} H_b\left(\frac{1}{k_i}, \ldots, \frac{1}{k_i}\right).$$

   Again, this makes sense. Suppose that $X$ is a random variable with values $\{x_1, x_2, \ldots, x_n\}$, and each $x_i$ is equally likely, with probability $\frac{1}{n}$. Partition $\{x_1, \ldots, x_n\}$ into $m$ sets $B_1, B_2, \ldots, B_m$ of cardinalities $k_1, k_2, \ldots, k_m$ respectively. Define the value of $Y$ to be equal to $j$ if the value of $X$ lies in $B_j$. Note that $P(Y = j) = \frac{k_j}{n}$, and $P(X = x_i | Y = j) = \left(\frac{k_j}{n}\right) / \left(\frac{k_i}{n}\right) = \frac{1}{k_i}$ if $x_i \in B_j$. Now
   
   $$H_b(Y) = H_b\left(\frac{k_1}{n}, \ldots, \frac{k_m}{n}\right)$$

   Assuming that the value of $X$ lies in $B_j$, each element of $B_j$ is equally likely with chance $\frac{1}{k_i}$. This adds an additional uncertainty of
   
   $$H_b\left(\frac{1}{k_1}, \ldots, \frac{1}{k_i}\right).$$

   The expression
   
   $$\sum_{i=1}^{m} \frac{k_i}{n} H_b\left(\frac{1}{k_1}, \ldots, \frac{1}{k_i}\right)$$

   can be thought of as the expected uncertainty of choosing $x_i$ within a set $B_j$. Also
   
   $$H_b(Y) = H_b\left(\frac{k_1}{n}, \ldots, \frac{k_m}{n}\right)$$

   can be thought of as the expected uncertainty of choosing a set $B_j$.

   It is proven in S. Roman’s book *Coding and Information Theory*, 1.1, that the properties 1,2 and 3 uniquely define $H_b$ (up to a scalar).
Proof.

1. This follows from the fact that $x \log_b(x)$ can be extended to a continuous function for all $x \geq 0$.

2. 
   
   \[ H_b\left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right) = -\frac{1}{n} \log_b\left(\frac{1}{n}\right) - \cdots - \frac{1}{n} \log_b\left(\frac{1}{n}\right) = - \log_b\left(\frac{1}{n}\right) = \log_b(n) \]

   and similarly $H_b\left(\frac{1}{n+1}, \ldots, \frac{1}{n+1}\right) = \log_b(n + 1)$. So indeed
   
   \[ H_b\left(\frac{1}{n}, \ldots, \frac{1}{n}\right) = \log_b(n) < \log_b(n + 1) = H_b\left(\frac{1}{n+1}, \ldots, \frac{1}{n+1}\right). \]

3. 
   
   \[ H_b\left(\frac{k_1}{n}, \ldots, \frac{k_m}{n}\right) = \sum_{i=1}^{m} -\frac{k_i}{n} \log_b\left(\frac{k_i}{n}\right) = \sum_{i=1}^{m} -\frac{k_i}{n} \log_b(k_i) + \frac{k_i}{n} \log_b(n) = \]
   
   \[ = \sum_{i=1}^{m} -\frac{k_i}{n} \log_b(k_i) + \sum_{i=1}^{m} \frac{k_i}{n} \log_b(n) = \sum_{i=1}^{m} -\frac{k_i}{n} \log_b(k_i) + \log_b(n) \]

   and
   
   \[ \sum_{i=1}^{m} \frac{k_i}{n} H\left(\frac{1}{k_i}, \ldots, \frac{1}{k_i}\right) = \sum_{i=1}^{m} \frac{k_i}{n} \log_b(k_i). \]

   So we have
   
   \[ H_b\left(\frac{1}{n}, \ldots, \frac{1}{n}\right) = \log_b(n) = H_b\left(\frac{k_1}{n}, \ldots, \frac{k_m}{n}\right) + \sum_{i=1}^{m} \frac{k_i}{n} H\left(\frac{1}{k_i}, \ldots, \frac{1}{k_i}\right). \]

4. **Further Properties**

   Before we prove some further properties, we will prove two lemmas:

   **Lemma 5.** Let $e = 2.71828 \cdots$ be the Euler number. We have the inequality
   
   \[ \log_e(x) \leq x - 1 \]
   
   for all $x > 0$ with equality if and only if $x = 1$.

   **Proof.** Let $f(x) = \log_e(x) - x + 1$. Then we have $f'(x) = \frac{1}{x} - 1$. So $f'(x) < 0$ if $x > 1$ and $f'(x) > 0$ if $0 < x < 1$. So $f$ is increasing for $0 < x < 1$ and $f$ is decreasing for $x > 1$ and moreover, $f(1) = 0$. We see that $f(x) \leq 0$ for all $x > 0$ with equality if $x = 1$. \qed

   **Lemma 6.** Suppose $b > 1$. Let $p_1, p_2, \ldots, p_n, q_1, \ldots, q_n$ be nonnegative real numbers with $\sum_{i=1}^{n} p_i = 1$ and $\sum_{i=1}^{n} q_i \leq 1$. Then
   
   \[ \sum_{i=1}^{n} p_i \log_b \frac{1}{p_i} \leq \sum_{i=1}^{n} p_i \log_b \frac{1}{q_i} \]

   Equality holds if $p_i = q_i$ for all $i$. 

Proof. It doesn’t matter which base \( b \) we take, as long as \( b > 1 \). Let us take \( b = e \). If \( p_i = 0 \) then we might as well omit \( p_i \) and \( q_i \). If \( q_i = 0 \) (and \( p_i \neq 0 \)) then the righthand side should be interpreted as \( \infty \) where the lefthand side is finite. Let us assume that all \( p_i \) and \( q_i \) are nonzero. Consider

\[
\sum_{i=1}^{n} p_i \log_e \frac{1}{p_i} - \sum_{i=1}^{n} p_i \frac{1}{q_i} = \sum_{i=1}^{n} p_i \log_e \frac{q_i}{p_i} = \sum_{i=1}^{n} p_i (q_i - 1) = \sum_{i=1}^{n} q_i - \sum_{i=1}^{n} p_i \leq 0
\]

using the previous lemma. Equality holds if \( p_i = q_i \) for all \( i \).

\[\square\]

Theorem 7. If \( X \) and \( Y \) are random variables with values \( \{x_1, \ldots, x_n\} \) and \( \{y_1, \ldots, y_m\} \) respectively, then

\[ H(X, Y) \leq H(X) + H(Y) \]

with equality if and only if \( X \) and \( Y \) are independent.

Proof. Put \( p_i = P(X = x_i) \), \( q_j = P(Y = y_j) \) let \( r_{i,j} = P(X = x_i, Y = y_j) \). Note that \( \sum_i r_{i,j} = q_j \) and \( \sum_j r_{i,j} = p_i \). Now

\[ H(X, Y) - H(X) - H(Y) = \sum_{i,j} r_{i,j} \log \frac{1}{r_{i,j}} - \sum_i p_i \log \frac{1}{p_i} - \sum_j q_j \log \frac{1}{q_j} = \]

\[ = \sum_{i,j} q_{i,j} \log \frac{1}{q_{i,j}} - q_{i,j} \log \frac{1}{p_i} - q_{i,j} \log \frac{1}{q_j} = \sum_{i,j} q_{i,j} \log \frac{1}{q_{i,j}} - \sum_{i,j} q_{i,j} \log \frac{1}{p_i q_j} \leq 0 \]

because of the previous lemma. Equality if and only if \( q_{i,j} = p_i q_j \), in other words, if and only if \( X \) and \( Y \) are independent.

\[\square\]

Proposition 8. If \( X \) is a random variable with \( n \) values \( \{x_1, x_2, \ldots, x_n\} \), then \( H(X) \leq \log(n) \) with equality if and only if \( P(X = x_i) = \frac{1}{n} \) for all \( i \).

Proof. The proof is left as an exercise.

\[\square\]

An other important property of entropy is the grouping axiom. This is a generalization of property 3 in the previous section.

Proposition 9. Suppose that \( q_i = p_{i,1} + \cdots + p_{i,k_i} \) for \( i = 1, \ldots, m \) with \( p_{i,j} \geq 0 \) for all \( i, j \), and \( q_1 + q_2 + \cdots + q_m = 1 \). Then we have the inequality

\[ H(p_{1,1}, p_{1,2}, \ldots, p_{m,k_m}) = H(q_1, q_2, \ldots, q_m) + \sum_{i=1}^{m} q_i H\left( \frac{p_{i,1}}{q_i}, \frac{p_{i,2}}{q_i}, \ldots, \frac{p_{i,k_i}}{q_i} \right) \]

Proof. Again, the proof is left as an exercise.

\[\square\]

Corollary 10. Suppose \( X \) and \( Y \) are random variables with values \( \{x_1, x_2, \ldots, x_n\} \) and \( \{y_1, \ldots, y_m\} \) respectively. Then we have

\[ H(X, Y) \geq H(X) \]

\[\square\]
Proof. Put \( p_{i,j} = P(X = x_i, Y = y_j) \), and \( q_i = P(X = x_i) = p_{i,1} + p_{i,2} + \ldots + p_{i,m} \) for \( i = 1, \ldots, n \). Then we get

\[
H(X, Y) = H(p_{1,1}, p_{1,2}, \ldots, p_{n,m}) = H(q_1, q_2, \ldots, q_n) + \sum_{i=1}^{n} q_i H\left(\frac{p_{i,1}}{q_i}, \ldots, \frac{p_{i,m}}{q_i}\right) \\
\geq H(q_1, \ldots, q_n) = H(X).
\]

\(\square\)

Definition 11. We make the following definitions

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

Here \( I(X, Y) \) is called the \textit{mutual information} of \( X \) and \( Y \). Note that the mutual information is equal to 0 if and only if \( X \) and \( Y \) are independent. In that case the output \( Y \) does not give any information of the output of \( X \) whatsoever. Notice that \( 0 \leq I(X, Y) \leq \min(H(X), H(Y)) \). One can think of \( H(X|Y) \) as the amount of information of \( X \) which is not shared by \( Y \). Notice that \( 0 \leq H(X|Y) \leq H(X) \) and \( H(X|Y) = H(X) - I(X, Y) \).

Lemma 12. Suppose that \( X, Y, Z \) are three random variables with values \( \{x_1, \ldots, x_n\} \), \( \{y_1, \ldots, y_m\} \) and \( \{z_1, z_2, \ldots, z_l\} \). If

\[
P(X = x_i, Y = y_j | Z = z_k) = P(X = x_i | Z = z_k)P(Y = y_j | Z = z_k)
\]

for all \( i, j, k \), then

\[
\]

Proof. Exercise. \(\square\)

5. Channel and Channel Capacity

Instead of always using the alphabet with only 0 and 1 we may also consider other alphabets. A code of length \( n \) and alphabet \( A = \{\alpha_1, \ldots, \alpha_s\} \) is a subset of \( C \) of \( A^n \). We considered before the binary channel which is a channel for which both the input and the output are in the alphabet \( \{0, 1\} \). We were also assuming that the channel was symmetric. This means the that probability that a 0 is changed to a 1 is the same as the probability that a 1 is changed into a 0. We will for a moment consider much more general channels. For one thing, we will not assume any longer that the input alphabet is the same as the output alphabet. We make the following definition.

Definition 13. A channel is an \textit{input alphabet} \( I = \{\alpha_1, \ldots, \alpha_s\} \) and \textit{output alphabet} \( O = \{\beta_1, \ldots, \beta_t\} \) together with conditional probabilities \( P(Y = \beta_i \mid X = \alpha_j) \geq 0 \).
Here $X$ is thought of as the random variable of the input, and $Y$ is thought of as the random variable of the output. We must have

$$\sum_{i=1}^{t} P(Y = \beta_i | X = \alpha_j) = 1$$

for all $j$.

We will now discuss the capacity of a channel. The capacity is roughly the maximum amount of information that can be sent over the channel if a single (random) letter of the alphabet is sent.

First we note that if a distribution for the input random variable $X$ is chosen (so we have chosen real numbers $P(X = x_i)$ for $i = 1, \ldots, s$), then the distribution of the output random variable $Y$ is determined, since we have the formula

$$P(Y = \beta_i) = \sum_{j=1}^{s} P(Y = \beta_i | X = \alpha_j)P(X = \alpha_j).$$

**Definition 14.** The capacity of a channel with input alphabet $I = \{\alpha_1, \ldots, \alpha_s\}$, output alphabet $O = \{\beta_1, \ldots, \beta_t\}$ and probabilities $P(X = \alpha_i | Y = \beta_j)$ is the maximum value of $I(X,Y)$, where the maximum is taken over all possible probability distributions of $X$.

A probability distribution of $X$ is a collection of nonnegative real numbers $p_i = P(X = \alpha_i)$ with

$$p_1 + p_2 + \cdots + p_s = 1.$$  

Note that the set of all such $(p_1, p_2, \ldots, p_s)$ is a compact set in $\mathbb{R}^t$ and since $I(X,Y)$ depends in a continuous way, $I(X,Y)$ will have a maximum value.

**Example 15.** We compute the capacity of a symmetric binary channel with error probability $p$. A binary channel has input alphabet $I = \{0,1\}$ and output alphabet $O = \{0,1\}$. Now

$$P(Y = 1 | X = 0) = P(Y = 0 | X = 1) = p$$

and

$$P(Y = 1 | X = 1) = P(Y = 0 | X = 0) = 1 - p$$

because the channel is symmetric with error probability $p$.

Consider an probability distribution of $X$, say

$$P(X = 0) = q, \quad \text{and } P(X = 1) = 1 - q.$$ 

Then

$$P(Y = 0, X = 0) = (1 - p)q, \quad P(Y = 1, X = 0) = pq,$$

$$P(Y = 0, X = 1) = p(1 - q), \quad P(Y = 1, X = 1) = (1 - p)(1 - q).$$
Using the grouping axiom, using \( q = pq + (1 - p)q \) and \( (1 - q) = (1 - q)p + (1 - q)(1 - p) \), we get
\[
H(X, Y) = H(X) + qH(p, (1 - p)) + (1 - q)H(p, (1 - p)) = H(X) + H(p, 1 - p).
\]

Now
\[
I(X, Y) = H(Y) + H(X) - H(X, Y) = H(Y) - H(p, 1 - p).
\]
Since \( Y \) has only 2 possible values, we have
\[
H(Y) \leq \log 2 = 1.
\]
So in particular
\[
I(X, Y) \leq 1 - H(p, 1 - p).
\]
On the other hand, if we choose \( P(X = 0) = P(X = 1) = \frac{1}{2} \), then it is easy to compute
\[
P(Y = 0) = P(Y = 0, X = 0) + P(Y = 0, X = 1) = \frac{1}{2}(1 - p) + \frac{1}{2}p = \frac{1}{2}
\]
and
\[
P(Y = 1) = \frac{1}{2}.
\]
In that case \( H(Y) = 1 \) and we get
\[
I(X, Y) = 1 - H(p, 1 - p).
\]
We have proven that the channel capacity if the symmetric binary channel with error probability \( p \) is equal to
\[
1 - H(p, 1 - p) = 1 + p \log p + (1 - p) \log(1 - p).
\]
Note that this is exactly the quanity which appeared in Shannon’s Theorem. Shannon’s Theorem said roughly that if the rate \( R \) is smaller than this channel capacity \( 1 + p \log p + (1 - p) \log(1 - p) \), then good codes with this rate and arbitrary small error probability exist.

6. A converse of Shannon’s Theorem

Suppose that \( C = \{x_1, x_2, \ldots, x_m\} \subseteq I^n \) is a code of length \( n \) in the alphabet \( I \). The code is sent over a channel with input alphabet \( I \) and output alphabet \( O \). There is a decoder which “guesses” the original codeword if a word \( y \in O^n \) is received. We can think of this decoder as some decision scheme, a function from \( f : O^n \to C \). Let \( X \) be the random variable of the code, and let \( Y \) be the random variable of the received word by the decoder. (So now the values of \( X \) and \( Y \) are words, rather than a single letter. We can think of \( X \) and \( Y \) as random vectors \( X = (X_1, X_2, \ldots, X_n) \) and \( Y = (Y_1, Y_2, \ldots, Y_n) \) where \( X_i \) and \( Y_i \) are the \( i \)-th letter of the values of \( X \) and \( Y \) respectively.)

**Theorem 16** (Fano’s inequality). We have the following inequality
\[
H(X \mid Y) \leq \log(p_c) + p_c \log(m - 1).
\]

**Proof.** The proof can be found in S. Roman, *Coding and Information Theory*, Theorem 3.3.5. □
We can think of $H(X \mid Y)$ as the information of $X$, the input, which is lost. The inequality says that if information is lost, meaning that $H(X \mid Y) > 0$, then the error probability $p_e$ cannot be arbitrarily small.

Using Fano’s inequality, we can relate the error probability and the channel capacity.

**Theorem 17** (Converse of Shannon’s Theorem). Suppose that $R$ is a real number which is larger than the capacity $C$. There exists a positive real number $\delta$ such that whenever $C \subset I^n$ is a code of length $m$ with $\log(m)/n > R$ and every word of $C$ appears with equal probability $1/m$, then the error probability of the code $C$ is at least $\delta$.

**Proof.** We have

$$H(X|Y) \leq H(p_e) + p_e \log(m-1).$$

Now

$$H(X) - I(X,Y) \leq H(p_e) + p_e \log(m-1).$$

Since all codewords have equal probability $1/m$, we have $H(X) = \log m$.

Note that

$$H(Y|X) = \sum_{i=1}^{n} H(Y_i|X)$$

since the errors are independent. Moreover,

$$H(X, Y_i) - H(X_i) = H((X, Y_i)|X_i) = H(X|X_i) + H(Y_i|X_i) = H(X) - H(X_i) + H(X_i, Y_i) - H(X_i).$$

so

$$H(Y_i|X) = H(Y_i, X) - H(X) = H(Y_i, X_i) - H(X_i) = H(Y_i|X_i).$$

since $Y_i$ depends only on $X_i$. We obtain

$$H(Y|X) = \sum_{i=1}^{n} H(Y_i|X_i)$$

and

$$I(Y, X) = H(Y) - H(Y|X) \leq \sum_{i=1}^{n} H(Y_i) - H(Y_i|X_i) = \sum_{i=1}^{n} I(Y_i, X_i) = nC$$

where $C$ is the channel capacity. We get

$$\log m - nC \leq H(p_e) + p_e \log(m-1).$$

Dividing by $n$ yields

$$R - C \leq \frac{\log(m)}{n} - C \leq \frac{H(p_e)}{n} + p_e \frac{\log(m-1)}{n} \leq \frac{H(p_e)}{n} + p_e R.$$

Let $\delta_n$ be the smallest possible value of $p_e$ for which the inequality

$$R - C \leq \frac{H(p_e)}{n} + p_e R.$$

holds. Clearly $\delta_n > 0$ for all $n$, and $\lim_{n \to \infty} \delta_n = R - C > 0$. Define $\delta$ as the infimum of all $\delta_n$. Then $\delta > 0$ and $p_e \geq \delta$ for any choice of $C$ as in the theorem. \qed