Information-Theoretic Modeling
Lecture 3: Source Coding: Theory

Teemu Roos
Department of Computer Science, University of Helsinki
Fall 2014
1. **Entropy and Information**
   - Entropy
   - Information Inequality
   - Data Processing Inequality

2. **Data Compression**
   - Asymptotic Equipartition Property (AEP)
   - Typical Sets
   - Noiseless Source Coding Theorem
Given a discrete random variable $X$ with pmf $p_X$, we can measure the amount of “surprise” associated with each outcome $x \in \mathcal{X}$ by the quantity

$$I_X(x) = \log_2 \frac{1}{p_X(x)}.$$ 

The less likely an outcome is, the more surprised we are to observe it. (The point in the log-scale will become clear shortly.)

The **entropy** of $X$ measures the expected amount of “surprise”:

$$H(X) = E[I_X(X)] = \sum_{x \in \mathcal{X}} p_X(x) \log_2 \frac{1}{p_X(x)}.$$
Binary Entropy Function

For binary-valued $X$, with $p = p_X(1) = 1 - p_X(0)$, we have

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

1. **the joint entropy** of two (or more) random variables:

\[
H(X, Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_{X,Y}(x, y) \log_2 \frac{1}{p_{X,Y}(x, y)} ,
\]

2. **the entropy of a conditional distribution**:

\[
H(X \mid Y = y) = \sum_{x \in \mathcal{X}} p_{X \mid Y}(x \mid y) \log_2 \frac{1}{p_{X \mid Y}(x \mid y)} ,
\]

3. **and the conditional entropy**:

\[
H(X \mid Y) = \sum_{y \in \mathcal{Y}} p(y) H(X \mid Y = y)
\]

\[
= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_{X,Y}(x, y) \log_2 \frac{1}{p_{X \mid Y}(x \mid y)} .
\]
The joint entropy $H(X, Y)$ measures the uncertainty about the pair $(X, Y)$.

The entropy of the conditional distribution $H(X \mid Y = y)$ measures the uncertainty about $X$ when we know that $Y = y$.

The conditional entropy $H(X \mid Y)$ measures the expected uncertainty about $X$ when the value $Y$ is known.
Chain Rule of Entropy

Remember the chain rule of probability:

\[ p_{X,Y}(x,y) = p_Y(y) \cdot p_{X|Y}(x \mid y) . \]

For the entropy we have:

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

**Proof.**

\[ p_{X,Y}(x,y) = p_Y(y) \cdot p_{X|Y}(x \mid y) \]

Next apply \( \log(ab) = \log a + \log b. \)
Chain Rule of Entropy

Remember the chain rule of probability:

\[ p_{X,Y}(x, y) = p_Y(y) \cdot p_{X|Y}(x | y) . \]

For the entropy we have:

Chain Rule of Entropy

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

Proof.

\[ \log_2 p_{X,Y}(x, y) = \log_2 p_Y(y) + \log_2 p_{X|Y}(x | y) \]

Next apply \( \log a = - \log(1/a) \).
Chain Rule of Entropy

Remember the chain rule of probability:

\[ p_{X,Y}(x, y) = p_Y(y) \cdot p_{X|Y}(x \mid y). \]

For the entropy we have:

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

**Proof.**

\[ \log_2 \frac{1}{p_{X,Y}(x, y)} = \log_2 \frac{1}{p_Y(y)} + \log_2 \frac{1}{p_{X|Y}(x \mid y)}. \]

\[ \iff \quad E \left[ \log_2 \frac{1}{p_{X,Y}(x, y)} \right] = E \left[ \log_2 \frac{1}{p_Y(y)} \right] + E \left[ \log_2 \frac{1}{p_{X|Y}(x \mid y)} \right]. \]

\[ \iff \quad H(X, Y) = H(Y) + H(X \mid Y). \]
Chain Rule of Entropy

Remember the chain rule of probability:

\[ p_{X,Y}(x,y) = p_Y(y) \cdot p_{X|Y}(x|y). \]

For the entropy we have:

Chain Rule of Entropy

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

The rule can be extended to more than two random variables:

\[ H(X_1,\ldots,X_n) = \sum_{i=1}^{n} H(X_i | H_1,\ldots,H_{i-1}). \]

\[ X \perp Y \iff H(X|Y) = H(X) \iff H(X,Y) = H(X) + H(Y). \]
Chain Rule of Entropy

Remember the chain rule of probability:

\[ p_{X,Y}(x,y) = p_Y(y) \cdot p_{X|Y}(x \mid y). \]

For the entropy we have:

**Chain Rule of Entropy**

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

The rule can be extended to more than two random variables:

\[ H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i \mid H_1, \ldots, H_{i-1}). \]

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

*Logarithmic* scale makes entropy **additive**.
Mutual Information

The mutual information

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

measures the average decrease in uncertainty about \( X \) when the value of \( Y \) becomes known.

Mutual information is symmetric (chain rule):

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

On the average, \( X \) gives as much information about \( Y \) as \( Y \) gives about \( X \).
Relationships between Entropies

\[ H(X,Y) \]
\[ H(X) \]
\[ H(Y) \]
\[ H(X \mid Y) \]
\[ I(X ; Y) \]
\[ H(Y \mid X) \]
The relative entropy or **Kullback-Leibler divergence** between (discrete) distributions $p_X$ and $q_X$ is defined as

$$D(p_X \parallel q_X) = \sum_{x \in \mathcal{X}} p_X(x) \log_2 \frac{p_X(x)}{q_X(x)} .$$

(We consider $p_X(x) \log_2 \frac{p_X(x)}{q_X(x)} = 0$ whenever $p_X(x) = 0$.)
Information Inequality

Kullback-Leibler Divergence

The relative entropy or **Kullback-Leibler divergence** between (discrete) distributions $p_X$ and $q_X$ is defined as

$$D(p_X \parallel q_X) = \sum_{x \in \mathcal{X}} p_X(x) \log_2 \frac{p_X(x)}{q_X(x)}.$$ 

Information Inequality

For any two (discrete) distributions $p_X$ and $q_X$, we have

$$D(p_X \parallel q_X) \geq 0$$

with equality iff $p_X(x) = q_X(x)$ for all $x \in \mathcal{X}$.

**Proof.** Gibbs!
Kullback-Leibler Divergence

The information inequality implies

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

Proof.

\[ I(X ; Y) = H(X) - H(X | Y) \]
\[ = H(X) + H(Y) - H(X, Y) \]
\[ = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_{X,Y}(x, y) \log_2 \frac{p_{X,Y}(x, y)}{p_X(x) p_Y(y)} \]
\[ = D(p_{X,Y} \| p_X p_Y) \geq 0. \]

In addition, \( D(p_{X,Y} \| p_X p_Y) = 0 \) iff \( p_{X,Y}(x, y) = p_X(x) p_Y(y) \) for all \( x \in \mathcal{X}, y \in \mathcal{Y} \). This means that variables \( X \) and \( Y \) are independent iff \( I(X ; Y) = 0 \).
Properties of entropy:

1. \( H(X) \geq 0 \)

   \textit{Proof.} \( p_X(x) \leq 1 \Rightarrow \log_2 \frac{1}{p_X(x)} \geq 0. \)

2. \( H(X) \leq \log_2 |\mathcal{X}| \)

   \textit{Proof.} Let \( u_X(x) = \frac{1}{|\mathcal{X}|} \) be the uniform distribution over \( \mathcal{X} \).

\[
0 \leq D(p_X \parallel u_X) = \sum_{x \in \mathcal{X}} p_X(x) \log_2 \frac{p_X(x)}{u_X(x)} = \log_2 |\mathcal{X}| - H(X).
\]
Properties of Entropy

Properties of entropy:

1. \( H(X) \geq 0 \)

   \textit{Proof.} \( p_X(x) \leq 1 \Rightarrow \log_2 \frac{1}{p_X(x)} \geq 0. \)

2. \( H(X) \leq \log_2 |\mathcal{X}| \)

A \textit{combinatorial} approach to the definition of information (Boltzmann, 1896; Hartley, 1928; Kolmogorov, 1965):

\[ S = k \ln W. \]
Ludvig Boltzmann (1844–1906)
Properties of Entropy

Properties of entropy:

1. \( H(X) \geq 0 \)

   *Proof.* \( p_X(x) \leq 1 \Rightarrow \log_2 \frac{1}{p_X(x)} \geq 0. \)

2. \( H(X) \leq \log_2 |\mathcal{X}| \)

   A *combinatorial* approach to the definition of information (Boltzmann, 1896; Hartley, 1928; Kolmogorov, 1965):
   \[ S = k \ln W . \]

3. \( H(X \mid Y) \leq H(X) \)

   *Proof.*
   \[ 0 \leq I(X ; Y) = H(X) - H(X \mid Y) . \]
Properties of Entropy

Properties of entropy:

1. \( H(X) \geq 0 \)

   Proof. \( p_X(x) \leq 1 \Rightarrow \log_2 \frac{1}{p_X(x)} \geq 0. \)

2. \( H(X) \leq \log_2 |X| \)

   A **combinatorial** approach to the definition of information (Boltzmann, 1896; Hartley, 1928; Kolmogorov, 1965):
   \[ S = k \ln W. \]

3. \( H(X \mid Y) \leq H(X) \)

   *On the average*, knowing another r.v. can only reduce uncertainty about \( X \). However, note that \( H(X \mid Y = y) \) may be greater than \( H(X) \) for some \( y \) — “contradicting evidence”.
Chain Rule of Mutual Information

The **conditional mutual information** of variables $X$ and $Y$ given $Z$ is defined as

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

Chain Rule of Mutual Information

For random variables $X$ and $Y_1, \ldots, Y_n$ we have

$$I(X; Y_1, \ldots, Y_n) = \sum_{i=1}^{n} I(X; Y_i | Y_1, \ldots, Y_{i-1}).$$

Independence among $Y_1, \ldots, Y_n$ implies

$$I(X; Y_1, \ldots, Y_n) = \sum_{i=1}^{n} I(X; Y_i).$$
Let $X, Y, Z$ be (discrete) random variables. If $Z$ is *conditionally independent of $X$ given $Y*,$ i.e., if we have

$$p_{Z|X,Y}(z | x, y) = p_{Z|Y}(z | y) \quad \text{for all } x, y, z,$$

then $X, Y, Z$ form a **Markov chain** $X \rightarrow Y \rightarrow Z$.

For instance, $Y$ is a “noisy” measurement of $X,$ and $Z = f(Y)$ is the outcome of deterministic data processing performed on $Y,$ then we have $X \rightarrow Y \rightarrow Z.$

This implies that

$$I(X ; Z | Y) = H(Z | Y) - H(Z | Y, X) = 0.$$  

When $Y$ is known, $Z$ doesn’t give any extra information about $X$ (and vice versa).
Data Processing Inequality

Assuming that $X \rightarrow Y \rightarrow Z$ is a Markov chain, we get

$$I(X ; Y, Z) = I(X ; Z) + I(X ; Y | Z)$$

$$= I(X ; Y) + I(X ; Z | Y) .$$

Now, because $I(X ; Z | Y) = 0$, and $I(X ; Y | Z) \geq 0$, we obtain:

**Data Processing Inequality**

If $X \rightarrow Y \rightarrow Z$ is a Markov chain, then we have

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

No data-processing can increase the amount of information that we have about $X$. 
1 Entropy and Information
   • Entropy
   • Information Inequality
   • Data Processing Inequality

2 Data Compression
   • Asymptotic Equipartition Property (AEP)
   • Typical Sets
   • Noiseless Source Coding Theorem
If $X_1, X_2, \ldots$ is a sequence of independent and identically distributed (i.i.d.) r.v.'s with domain $\mathcal{X}$ and pmf $p_X$, then

$$\log_2 \frac{1}{p_X(X_1)}, \log_2 \frac{1}{p_X(X_2)}, \ldots$$

is also an i.i.d. sequence of r.v.'s.

The expected values of the elements of the above sequence are all equal to the entropy:

$$E \left[ \log_2 \frac{1}{p_X(X_i)} \right] = \sum_{x \in \mathcal{X}} p_X(x) \log_2 \frac{1}{p_X(x)} = H(X) \quad \text{for all } i \in \mathbb{N}.$$
The i.i.d. assumption is equivalent to

\[ p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i) . \]
The i.i.d. assumption is equivalent to

$$\frac{1}{p(x_1, \ldots, x_n)} = \prod_{i=1}^{n} \frac{1}{p_X(x_i)}.$$
The i.i.d. assumption is equivalent to

\[
\log_2 \frac{1}{p(x_1, \ldots, x_n)} = \log_2 \prod_{i=1}^{n} \frac{1}{p_X(x_i)}. 
\]
The i.i.d. assumption is equivalent to

\[ \log_2 \frac{1}{p(x_1, \ldots, x_n)} = \sum_{i=1}^{n} \log_2 \frac{1}{p_X(x_i)} . \]
The i.i.d. assumption is equivalent to

\[
\frac{1}{n} \log_2 \frac{1}{p(x_1, \ldots, x_n)} = \frac{1}{n} \sum_{i=1}^{n} \log_2 \frac{1}{p(x_i)}.
\]
The i.i.d. assumption is equivalent to

\[
\frac{1}{n} \log_2 \frac{1}{p(x_1, \ldots, x_n)} = \frac{1}{n} \sum_{i=1}^{n} \log_2 \frac{1}{p_X(x_i)}.
\]

By the (weak) law of large numbers, the average on the right-hand side converges in probability to its mean, i.e., the entropy:

\[
\lim_{n \to \infty} \Pr \left[ \left| \frac{1}{n} \sum_{i=1}^{n} \log_2 \frac{1}{p_X(X_i)} - H(X) \right| < \epsilon \right] = 1 \quad \text{for all } \epsilon > 0.
\]
The i.i.d. assumption is equivalent to

\[
\frac{1}{n} \log_2 \frac{1}{p(x_1, \ldots, x_n)} = \frac{1}{n} \sum_{i=1}^{n} \log_2 \frac{1}{p_X(x_i)}.
\]

Asymptotic Equipartition Property (AEP)

For i.i.d. sequences, we have

\[
\lim_{n \to \infty} \Pr \left[ \left| \frac{1}{n} \log_2 \frac{1}{p(x_1, \ldots, x_n)} - H(X) \right| < \epsilon \right] = 1
\]

for all \( \epsilon > 0 \).
The AEP states that for any $\epsilon > 0$, and large enough $n$, we have

\[
\Pr \left[ \left| \frac{1}{n} \log_2 \frac{1}{p(x_1, \ldots, x_n)} - H(X) \right| < \epsilon \right] \approx 1
\]

\[
H(X) - \epsilon < \frac{1}{n} \log_2 \frac{1}{p(x_1, \ldots, x_n)} < H(X) + \epsilon
\]
The AEP states that for any $\epsilon > 0$, and large enough $n$, we have

$$\Pr \left[ \left| \frac{1}{n} \log_2 \frac{1}{p(x_1, \ldots, x_n)} - H(X) \right| < \epsilon \right] \approx 1$$

$$n(H(X) - \epsilon) < \log_2 \frac{1}{p(x_1, \ldots, x_n)} < n(H(X) + \epsilon)$$
The AEP states that for any $\epsilon > 0$, and large enough $n$, we have

$$\Pr \left[ \left| \frac{1}{n} \log_2 \frac{1}{p(x_1, \ldots, x_n)} - H(X) \right| < \epsilon \right] \approx 1$$

$$2^n(H(X) - \epsilon) < \frac{1}{p(x_1, \ldots, x_n)} < 2^n(H(X) + \epsilon)$$
The AEP states that for any $\epsilon > 0$, and large enough $n$, we have

$$\Pr \left[ \left| \frac{1}{n} \log_2 \frac{1}{p(x_1, \ldots, x_n)} - H(X) \right| < \epsilon \right] \approx 1$$

$$2^{-n(H(X)+\epsilon)} < p(x_1, \ldots, x_n) < 2^{-n(H(X)-\epsilon)}$$

$$\Leftrightarrow \Pr \left[ p(x_1, \ldots, x_n) = 2^{-n(H(X)\pm \epsilon)} \right] \approx 1 .$$

Asymptotic Equipartition Property (informally)

“Almost all sequences are almost equally likely.”
Technically, the key step in the proof was using the weak law of large numbers to deduce

$$\lim_{n \to \infty} \Pr \left[ \left| \frac{1}{n} \sum_{i=1}^{n} \log_2 \frac{1}{p_X(X_i)} - H(X) \right| < \epsilon \right] = 1 \quad \text{for all } \epsilon > 0.$$ 

In other words, with high probability the average “surprisingness” \( \log_2 p_X(X_i) \) over the sequence is close to its expectation.
Typical Sets

The **typical set** \( A_{\varepsilon}^{(n)} \) is the set of sequences \((x_1, \ldots, x_n) \in \mathcal{X}^n\) with the property:

\[
2^{-n(H(X)+\varepsilon)} \leq p(x_1, \ldots, x_n) \leq 2^{-n(H(X)-\varepsilon)}.
\]

The AEP states that

\[
\lim_{n \to \infty} \Pr \left[ X^n \in A_{\varepsilon}^{(n)} \right] = 1.
\]

In particular, for any \( \varepsilon > 0 \), and large enough \( n \), we have

\[
\Pr \left[ X^n \in A_{\varepsilon}^{(n)} \right] > 1 - \varepsilon.
\]
Typical Sets

How many sequences are there in the typical set $A^{(n)}_{\epsilon}$?

We can use the fact that by definition each sequence has probability at least $2^{-n(H(X)+\epsilon)}$.

Since the total probability of all the sequences in $A^{(n)}_{\epsilon}$ is trivially at most 1, there can’t be too many of them.

$$1 \geq \sum_{(x_1, \ldots, x_n) \in A^{(n)}_{\epsilon}} p(x_1, \ldots, x_n)$$

$$\geq \sum_{(x_1, \ldots, x_n) \in A^{(n)}_{\epsilon}} 2^{-n(H(X)+\epsilon)} = 2^{-n(H(X)+\epsilon)} \left| A^{(n)}_{\epsilon} \right|$$

$$\Leftrightarrow \left| A^{(n)}_{\epsilon} \right| \leq 2^{n(H(X)+\epsilon)}.$$
Typical Sets

Is it possible that the typical set $A^{(n)}_{\epsilon}$ is very small?

This time we can use the fact that by definition each sequence has probability at most $2^{-n(H(X)-\epsilon)}$.

Since for large enough $n$, the total probability of all the sequences in $A^{(n)}_{\epsilon}$ is (by the AEP) at least $1 - \epsilon$, there can’t be too few of them.

$$1 - \epsilon < \Pr \left[ X^n \in A^{(n)}_{\epsilon} \right]$$

$$\leq \sum_{(x_1, \ldots, x_n) \in A^{(n)}_{\epsilon}} 2^{-n(H(X)-\epsilon)} = 2^{-n(H(X)-\epsilon)} |A^{(n)}_{\epsilon}|$$

$$\Leftrightarrow |A^{(n)}_{\epsilon}| > (1 - \epsilon)2^{n(H(X)-\epsilon)}.$$
Typical Sets

So the AEP guarantees that for small $\epsilon$ and large $n$:

1. The typical set $A^{(n)}_\epsilon$ has high probability.
2. The number of elements in the typical set is about $2^{nH(X)}$.

So what?
Typical Sets

So the AEP guarantees that for small \( \epsilon \) and large \( n \):

1. The typical set \( A_{\epsilon}^{(n)} \) has high probability.
2. The number of elements in the typical set is about \( 2^{nH(X)} \).

The number of all possible sequences \( (x_1, \ldots, x_n) \in \mathcal{X}^n \) of length \( n \) is \( |\mathcal{X}|^n \).

The maximum of entropy is \( \log_2 |\mathcal{X}| \). If \( H(X) = \log_2 |\mathcal{X}| \), we obtain

\[
\left| A_{\epsilon}^{(n)} \right| \approx 2^{nH(X)} = 2^n \log_2 |\mathcal{X}| = |\mathcal{X}|^n,
\]

i.e., the typical set can be as large as the whole set \( \mathcal{X}^n \).
Typical Sets

So the AEP guarantees that for small $\varepsilon$ and large $n$:

1. The typical set $A_\varepsilon^{(n)}$ has high probability.
2. The number of elements in the typical set is about $2^{nH(X)}$.

The number of all possible sequences $(x_1, \ldots, x_n) \in \mathcal{X}^n$ of length $n$ is $|\mathcal{X}|^n$.

However, for $H(X) < \log_2 |\mathcal{X}|$, the number of sequences in $A_\varepsilon^{(n)}$ is exponentially smaller than $|\mathcal{X}|^n$:

$$\frac{2^{nH(X)}}{2^{n\log_2 |\mathcal{X}|}} = 2^{-n\delta} \xrightarrow{n \to \infty} 0,$$

where $\delta = \log_2 |\mathcal{X}| - H(X) > 0$. 

Teemu Roos
Information-Theoretic Modeling
Typical Sets

$\chi^n: |\chi|^n$ elements

Typical Set

$A_\varepsilon^{(n)}: 2^{nH(X)}$ elements

A (relatively) small set that contains most of the probability mass.
Typical Sets

A (relatively) small set that contains most of the probability mass.
Typical Sets

\[ \chi^n : |\chi|^n \text{ elements} \]

A (relatively) small set that contains most of the probability mass.

Teemu Roos  Information-Theoretic Modeling
Typical Sets

\[ X^n : |X|^n \text{ elements} \]

A (relatively) small set that contains most of the probability mass.
A (relatively) small set that contains most of the probability mass.
Typical Sets

$\mathcal{X}^n: |\mathcal{X}|^n$ elements

A (relatively) small set that contains most of the probability mass.
Examples

If the source consists of i.i.d. bits $\mathcal{X} = \{0, 1\}$ with $p = p_X(1) = 1 - p_X(0)$, then we have

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p_X(x_i) = p^{\sum x_i} (1 - p)^{n - \sum x_i},$$

where $\sum x_i$ is the number of 1’s in $x^n$.

In this case, the typical set $A^{(n)}_{\epsilon}$ consists of sequences for which $\sum x_i$ is close to $np$. For such strings, we have

$$\log_2 \frac{1}{p(x_1, \ldots, x_n)} \approx \log_2 \frac{1}{p^{np}(1 - p)^{n(1-p)}}$$

$$= n \left( p \log_2 \frac{1}{p} + (1 - p) \log_2 \frac{1}{1 - p} \right) = nH(X).$$
If the source consists of i.i.d. rolls of a die $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$ with $p_j = p_{\mathcal{X}}(j), \ j \in \mathcal{X}$, then we have

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p_{\mathcal{X}}(x_i) = \prod_{j=1}^{6} p_j^{k_j},$$

where $k_j$ is the number of times $x_i = j$ in $x^n$.

In this case, the typical set $A_{\epsilon(n)}$ consists of sequences for which $k_j$ is close to $np_j$ for all $j \in \{1, 2, 3, 4, 5, 6\}$. For such strings, we have

$$\log_2 \frac{1}{p(x_1, \ldots, x_n)} \approx \log_2 \frac{1}{\prod_{j=1}^{6} p_j^{np_j}}$$

$$= n \left( \sum_{j=1}^{6} p_j \log \frac{1}{p_j} \right) = nH(X).$$
The AEP Code

We now construct a code from source strings \((x_1, \ldots, x_n) \in \mathcal{X}^n\) to binary sequences \(\{0, 1\}^*\) of arbitrary length.

Let \(x^n \in \mathcal{X}^n\) denote the sequence \((x_1, \ldots, x_n)\), and let \(\ell(x^n)\) denote the length (bits) of the codeword assigned to sequence \(x^n\).

The code we will construct has expected per-symbol codeword length arbitrarily close to the entropy

\[
E \left[ \frac{1}{n} \ell(x^n) \right] \leq H(X) + \epsilon ,
\]

for large enough \(n\).

This is the best achievable rate for uniquely decodable codes.
The AEP Code

We treat separately two kinds of source strings $x^n \in \mathcal{X}^n$:

1. the **typical** strings $x^n \in A_\epsilon^{(n)}$, and
2. the **non-typical** strings $x^n \in \mathcal{X}^n \setminus A_\epsilon^{(n)}$.

There are at most $2^{n(H(X)+\epsilon)}$ strings of the first kind. Hence, we can encode them using binary strings of length $n(H(X)+\epsilon) + 1$.

There are at most $|\mathcal{X}|^n$ strings of the second kind. Hence we can encode them using binary strings of length $n \log_2 |\mathcal{X}| + 1$.

Since the decoder must be able to tell which kind of a string it is decoding, we prefix the code by a 0 if $x^n \in A_\epsilon^{(n)}$ or by 1 if not. This adds one more bit in either case.
The AEP Code

To see what’s going on, consider the situation $H(X) < \log_2 |\mathcal{X}|$. This is the interesting case in which the code actually does result in compression.

In the first lecture we saw that any attempt to compress everything will fail because there are not enough short codewords.

We bypass this by splitting into two cases.

1. **Typical** strings are actually compressed. There are not too many of them, so there are enough short codewords.

2. **Non-typical** strings are not compressed. Because their total probability is low (AEP), this does not matter too much.
Expected Codelength of the AEP Code

Let us calculate the expected per-symbol codeword length:

\[
E[\ell(X^n)] = E \left[ \ell(X^n) \mid X^n \in A^{(n)}_\epsilon \right] \Pr \left[ X^n \in A^{(n)}_\epsilon \right] \\
+ E \left[ \ell(X^n) \mid X^n \notin A^{(n)}_\epsilon \right] \Pr \left[ X^n \notin A^{(n)}_\epsilon \right]
\]

\[
= (n(H(X) + \epsilon) + 2) \Pr \left[ X^n \in A^{(n)}_\epsilon \right] \\
+ (n \log_2 |\mathcal{X}| + 2) \Pr \left[ X^n \notin A^{(n)}_\epsilon \right]
\]

\[
\leq n(H(X) + \epsilon) + n \log |\mathcal{X}| \epsilon + 2 \quad \text{(AEP)}
\]

\[
= n(H(X) + \epsilon') ,
\]

where \( \epsilon' = \epsilon + \epsilon \log_2 |\mathcal{X}| + \frac{2}{n} \) can be made arbitrarily small by choosing \( \epsilon > 0 \) small enough, and letting \( n \) become large enough.
Optimality of the AEP Code

Dividing this bound by \( n \) gives the expected per-symbol codeword length of the “AEP code”:

\[
E \left[ \frac{1}{n} \ell(X^n) \right] \leq H(X) + \epsilon
\]

for any \( \epsilon > 0 \) and \( n \) large enough.

Optimality: By AEP, there are about \( 2^{nH(X)} \) sequences that have probability about \( 2^{-nH(X)} \). We can assign a codeword shorter than \( n(H(X) - \delta) \) to only a proportion of less than \( 2^{-n\delta} \) of these sequences (by a counting argument), and hence the expected per-symbol codeword length must be about \( H(X) \) or more.
Noiseless Source Coding Theorem

These two statements give the

9. THE FUNDAMENTAL THEOREM FOR A NOISELESS CHANNEL

We will now justify our interpretation of $H$ as the rate of generating information by proving that $H$ determines the channel capacity required with most efficient coding.

Theorem 9: Let a source have entropy $H$ (bits per symbol) and a channel have a capacity $C$ (bits per second). Then it is possible to encode the output of the source in such a way as to transmit at the average rate $\frac{C}{H} - \epsilon$ symbols per second over the channel where $\epsilon$ is arbitrarily small. It is not possible to transmit at an average rate greater than $\frac{C}{H}$.

(Shannon, 1948)

In the noiseless setting with binary code alphabet, the channel capacity is $C = \log_2 |\{0, 1\}| = 1$.

The theorem says that the achievable rates are given by

$$R = \lim_{n \to \infty} \frac{n}{\ell(x^n)} < \frac{1}{H(X)}.$$
Coming next

Next on the course:

1. brief excursion into noisy channel coding
2. source coding in practice: efficient algorithms.