Quantum Information Theory

Lectures by Sergii Strelchuk

Lecture 1

Definition: (Surprisal) Consider an event described by a random variable X which takes values $x \in J$ (J a discrete, finite, we call it the *alphabet*), and each x occurs with probability $p(x)$ for all $x \in J$. We will denote this as $X \sim p(x), x \in J$.

A measure of uncertainty given an outcome x is given by ***surprisal***:

$$\mathcal{I}(x) \coloneqq -\log p(x)$$

We will consider logarithms to be base 2, as we are working (temporarily) in the realm of classical information theory. Note that surprisal does not depend on the values x, but their probabilities.

Definition: (Shannon Entropy) The ***Shannon entropy*** $H(X)$ of a discrete random variable $X \sim p(x), x \in J$ is defined as:

$$H(X) \coloneqq -\sum_{x \in J} p(x) \log p(x)$$

Example (binary entropy): Let X be Bernoulli with probability p.

Then $H(X) = -p \log p - (1-p)\log(1-p) \coloneqq h(p)$.

Claude Shannon asked (and attempted to answer) a few questions, back in the day:

Question 1: What is the limit to which information can be reliably compressed in a manner such that it can be recovered "reliably" (*i.e.* with low probability of error, made specific later)?

This is often of interest, because we often have a physical limit (*e.g.* laptop hard drive) for storing information, and we will want to store as much as possible safely.

Question 2: What is the maximum amount of information that can be reliably transmitted per use of communication channel?

In other words, how do we store and how do we send information?

Answer 1: Shannon's Source Coding Theorem: the limit of reliable compression is given by the Shannon entropy of the 'source'.

We will consider the sources to be discrete, with a finite alphabet. One of the simplest examples of a source is a ***Memoryless Source***.

Definition: (Memoryless Source) This is characterised by $\{p(u)\}_{u \in J}$, where u is a 'letter', emitted with probability $p(u)$.

The source can emit multiple signals, each of which is given by a random variable U_i, characterised by $p(u) = \mathbb{P}[U_i = u], u \in J$ for $1 \leq i \leq n$.

So the source emits a sequence $(u_1, \ldots, u_n) \in J^n$, with probability

$$\begin{aligned} p(u_1, \ldots, u_n) &\coloneqq \mathbb{P}(U_1 = u_1, \ldots, U_n = u_n), u_i \in J \\ &= p(u_1) \ldots p(u_n) \end{aligned}$$

This is memoryless because the U_i are iid; this is an 'iid information source'. We characterise the Shannon entropy of the source as:

$$\begin{aligned} H(U) \equiv H(\{p(u)\}) &\coloneqq H(U_1) = H(U_2) = \cdots = H(U_n) \\ &= -\sum_{u \in J} p(u) \log p(u) \end{aligned}$$

Why is data compression even possible? As it happens, there is a lot of redundancy in data, for instance if some things occur more often than others; we can take advantage of this using:

Definition: (Variable/Fixed Length Coding) ***Variable length coding*** is a protocol by which more frequent outputs of the source are assigned shorter descriptions.

Fixed length coding is a protocol by which signals with a high probability of occurence are assigned unique fixed length binary strings. Signals with low probability of occurence are mapped to some fixed string *i.e.* we will not care about them.

We will see a brief example of variable length coding, but after that the remainder of the course will be focused on fixed length coding.

Example (Variable Length Coding): Suppose we have a source emitting one of eight signals in [8]. The table shows the probailities of each occuring, as well as a naive encoding C and a variable length encoding $\tilde{C}$.

u	1	2	3	4	5	6	7	8
$p(u)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{64}$	$\frac{1}{64}$	$\frac{1}{64}$	$\frac{1}{64}$
$C(u)$	000	001	010	011	100	101	110	111
$\tilde{C}(u)$	0	10	110	1110	111100	111101	111110	111111

While $\tilde{C}$ might seem counterintuitive (the strings are longer), this actually saves on the average length of the source code; for C, the average is 3, whereas for $\tilde{C}$ the average is 2.

As it happens, the average description here is equal to the entropy:

$$H(p_u) = -\sum_{u=1}^{8} p(u) \log p(u) = 2$$

We will see later that this is no coincidence.

In order to poperly reason about entropy and compression, we need to make a few definitons.

0 Classical Data Compression

Definition: (Classical Data Compression) Consider a memoryless source characterised by a sequence of random variables $U_1, \ldots U_n$, $p(u) = \mathbb{P}(U_k = u), u \in J$, $1 \leq k \leq n$, $|J| = d$. Let $U \sim p(u), u \in J$.

A ***compression map*** C^n ***of rate*** R maps each signal as follows:

$$C^n : \underline{u}^{(n)} = (u_1, \ldots, u_n) \mapsto \underline{x} = (x_1, \ldots, x_{nR})$$

where $u_i \in J$ and $x_k \in \{0, 1\}, 1 \leq k \leq \lceil nR \rceil$.

So C^n maps a message $\underline{u}^{(n)}$ into a codeword $\underline{x}$ of length nR.

A ***decompression map*** D^n goes the other way:

$$D^n : \underline{x} \in \{0,1\}^{\lceil nR \rceil} \mapsto \underline{u}'^{(n)} = (u'_1, \ldots, u'_n)$$

with probability $p(\underline{u}'^{(n)}|\underline{x})$.

So $D^n(C^n(\underline{u}^{(n)})) = \underline{u}'^{(n)}$.

The triple $C_n \coloneqq (C^n, D^n, R)$ is called a ***code*** of rate R and block length n.

The average probability of error of a code is given by:

$$P_{\text{av}}^{(n)}(C_n) \coloneqq \sum_{\underline{u}^{(n)} \in J^n} p(\underline{u}^{(n)}) \cdot \mathbb{P}(D^n(C^n(u^{(n)})) \neq u^{(n)})$$

Lecture 2

Recall that for a ***memoryless source***, $p(\underline{u}^{(n)}) = \prod_{i=1}^n p(u_i)$.

We say that a compression/decompression scheme is ***reliable*** if there exists a sequence of codes $\mathcal{C}_n$ such that:

$$\lim_{n\to\infty} P_{\text{av}}^{(n)}(\mathcal{C}_n) = 0$$

Equivalently, for any given $\varepsilon \in (0,1)$ and n large enough, we have that:

$$\sum_{\underline{u}^{(n)} \in J^n} P(\underline{u}^{(n)}) \mathbb{P}(D^n(C^n(\underline{u}^{(n)})) \neq \underline{u}^{(n)}) \leq \varepsilon$$

The ***data compression limit*** is given by

$$\inf\{R : \exists \mathcal{C}_n \coloneqq (C^n, D^n, R) \text{ s.t. } \lim_{n\to\infty} P_{\text{av}}^{(n)}(\mathcal{C}_n) = 0\}$$

Typical Sequences

Definition: (Typical Set) Consider an i.i.d information source described by random variables $U_1, \ldots, U_n$, $U_i \sim p(u), u \in J$. For any $\varepsilon \in (0,1)$, the ***typical set*** $T_\varepsilon^{(n)}$ is the set of sequences $(u_1, \ldots, u_n) \in J^n$ such that:

$$2^{-n(h(U)+\varepsilon)} \leq p(u_1, \ldots, u_n) \leq 2^{-n(H(U)-\varepsilon)}$$

where $H(U)$ is the Shannon entropy.

$|T_\varepsilon^{(n)}|$ denotes the size of the typical set, and $\mathbb{P}(T_\varepsilon^{(n)}$ the probability of a typical set.

The typical sequences in the set occur with roughly the same probability. We will see that for $\underline{u} \in J^n$, if $\underline{u} \in T_\varepsilon^{(n)}$ then $p(\underline{u}) \approx 2^{-nH(u)}$.

Is this a good way of capturing typicality? Consider the following situation.

We have a memoryless source that emits the letter $u \sim p(u)$. A typical sequence of length n will have length $np(u)$ on average.

The probability of such a sequence of n symbols is approximately given by

$$\begin{aligned}\prod_{u\in J} p(u)^{np(u)} &= \prod_{u\in J} 2^{\log p(u)^{np(u)}} \\ &= 2^{n\sum_{n\in J} p(u)\log p(u)} \\ &= 2^{-nH(U)}\end{aligned}$$

Theorem 0.1: (Typical Sequence Theorem) *Fix $\varepsilon \in (0,1)$. Then for any $\delta > 0$ there exists $n_0(\delta)$ such that the following is true (for any $n > n_0(\delta)$?):*

1. *If $(u_1, \ldots, u_n) \in T_\varepsilon^{(n)}$, then*

$$H(u) - \varepsilon \leq -\frac{1}{n} \log p(u_1, \ldots, u_n) \leq H(u) + \varepsilon$$

2.
$$\mathbb{P}(T_\varepsilon^{(n)}) > 1 - \delta$$

3.
$$|T_\varepsilon^{(n)}| \leq 2^{n(H(u)+\varepsilon)}$$

4.
$$|T_\varepsilon^{(n)}| > (1-\delta)2^{n(H(u)-\varepsilon)}$$

Corollary 0.2: *For any $\delta > 0$ there exists $n_0(\delta) > 0$ such that for any $n > n_0(\delta)$ the set J^n decomposes into disjoint subsets $A_\varepsilon^{(n)}, T_\varepsilon^{(n)}$ such that:*

1.
$$\mathbb{P}(A_\varepsilon^{(n)}) < \delta$$

2. *The probability of a typical sequence is bounded by*

$$2^{-n(H(u)+\varepsilon)} \leq \mathbb{P}(\underline{U}^{(n)} = \underline{u}^{(n)}) \leq 2^{-n(H(u)-\varepsilon)}$$

where $\underline{U}^{(n)} \coloneqq (U_1, U_2, \ldots, U_n)$ is a sequence of sources, $\underline{u}^{(n)} = (u_1, \ldots, u_n) \in J^n$

Theorem 0.3: (Shannon's Source Coding Theorem) *Suppose $\{U_i\}, U_i \sim p(u)$ is an i.i.d. information source, with Shannon entropy $H(u)$.*

Now suppose $R > H(u)$. Then there exists a reliable compression scheme of rate R for the given source.

Conversely, if $R < H(u)$ then ther is no reliable compression scheme of rate R.

Proof. Case: Suppose $R > H(u)$.

Choose $\varepsilon \in (0,1)$ such that $H(u) + \varepsilon < R$. Take n large enough so that $T_\varepsilon^{(n)}$ satisfies the Typical Sequence Theorem. Then for any $\delta > 0$ there are at most $2^{n(H(u)+\varepsilon)} < 2^{nR}$ ε-typical sequences, *i.e.* $|T_\varepsilon^{(n)}| < 2^{nR}$.

1) Divide all sequences J^n into 2 sets, $T_\varepsilon^{(n)}, A_\varepsilon^{(n)}$.

2) Order all the elements in $T_\varepsilon^{(n)}$ (say, lexicographically). Each ε-typical sequence can then be identified with its index.

   Since $|T_\varepsilon^{(n)}| \equiv \#$ typical sequences $\leq 2^{n(H(u)+\varepsilon)} < 2^{nR}$, the indexing requires no more than $\lceil nR \rceil$ bits.

   Examine each output of the source.

3) If it is typical, store the index (bit sequence) of a (the?) sequence. Prefix the typical bit seqeunces by 1. Then total length is $\lceil nR \rceil + 1$.

4) If the string is not typical, we will assign a fixed bit string $(00\ldots0)$ of length $\lceil nR \rceil + 1$.

The errors in the decompression then occur when trying to decompress the atypical sequences, which occur with vanishing probability for n large. So this is reliable.

□

Case: Suppose $R < H(u)$. Then the compression scheme is not reliable; we need a lemma.

Lemma 0.4: *Let $S(n)$ be a collection of strings $\underline{u}^{(n)}$ of length n of size $|S(n)| \leq 2^{nR}$, where $R < H(u)$. Each $\underline{u}^{(n)}$ is produced by the source with probability $p(\underline{u}^{(n)})$. Then for any $\delta > 0$ and sufficiently large n:*

$$\sum_{\underline{u}^{(n)} \in S(n)} p(\underline{u}^{(n)}) \leq \delta$$

Proof. (sketch). Split $S(n)$ into typical and atypical sequences. # typical sequences of $S(n)$ is $\leq$ the total # of sequences of $S(n)$. Hence $|S(n)| \leq 2^{nR}$. So each typical sequence has probability $\approx 2^{-nH(u)}$.

So the total probability of the typical sequences in $S(n)$ scales as $2^{nR} \cdot 2^{-nH(u)} = 2^{-n(h(u)-R)} \to 0$ as $n \to \infty$. More rigorously:

$$\begin{aligned}
\mathbb{P}(S(n)) &= \sum_{\underline{u}^{(n)} \in S(n)} p(\underline{u}^{(n)}) \\
&= \sum_{\underline{u}^{(n)} \in S(n) \cap T_\varepsilon^{(n)}} p(\underline{u}^{(n)}) + \sum_{\underline{u}^{(n)} \in S(n) \cap A_\varepsilon^{(n)}} P(\underline{u}^{(n)}) \\
&\leq |S(n)| 2^{-n(H(u)-\varepsilon)} + \mathbb{P}(A_\varepsilon^{(n)})
\end{aligned}$$

and both of these terms vanish in the limit due to the above discussion. □

Essentially this says that if we try to code with a rate too low, the size of the typical sets vanish at an exponential rate, and the compression becomes unreliable. Lecture 3