1 Asymptotic Equipartition Property

The following theorem is a consequence of the Weak Law of Large Numbers which will be crucial in what follows.

**Theorem 1.** ASYMPTOTIC EQUIPARTITION PROPERTY. Consider a sequence $X_1, X_2, \ldots, X_n$ of i.i.d. random variables with finite range distributed accordingly to a probability mass function $p$; then:

$$-\frac{1}{n} \log p(X_1, \ldots, X_n) \xrightarrow{p} H(p)$$

in words: the random variable $-\frac{1}{n} \log p(X_1, \ldots, X_n)$ converges in probability to the entropy $H(p)$.

**Proof.** Consider the new random variables $Y_1, Y_2, \ldots, Y_n$ defined by $Y_i = -\log p(X_i)$. Since the $X_i$ are i.i.d., then the $Y_i$ are i.i.d. too, given that functions of independent random variables are also independent random variables. The Weak Law of Large Numbers thus ensures that (2) holds.

$$\frac{1}{n} \sum_{i=1}^{n} Y_i \xrightarrow{p} \mathbb{E}_p[Y]$$

Note that:

$$-\frac{1}{n} \log p(X_1, \ldots, X_n) = -\frac{1}{n} \log \prod_{i=1}^{n} p(X_i) = -\frac{1}{n} \sum_{i=1}^{n} \log p(X_i) = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

$$H(p) = -\sum_{x \in \text{Im}(X)} p(x) \log p(x) = \mathbb{E}_p[Y]$$

Thus, claim (1) thus immediately follows by replacing (3) and (4) in (2). \qed

2 Typical sets

**Definition 1.** Consider a probability distribution $p(x)$ over a finite set $X$ and arbitrary $\epsilon > 0$ and $n \in \mathbb{N}$; the set $A^n_\epsilon(p)$ defined as follows:

$$A^n_\epsilon(p) = \left\{ x = (x_1, \ldots, x_n) \in X^n \mid \frac{1}{2^n[H(p)+\epsilon]} \leq p(x_1, \ldots, x_n) \leq \frac{1}{2^n[H(p)-\epsilon]} \right\}$$

is called the typical set for $p$ corresponding to $\epsilon$ and $n$. We will often write just $A^n_\epsilon$ instead of $A^n_\epsilon(p)$, when no confusion arises.

**Theorem 2.** The following properties hold:

1. $A^n_\epsilon(p) = \left\{ x \in X^n \mid -\frac{1}{n} \log p(x) - H(p) > \epsilon \right\}$.

2. $\mathbb{P}_p(A^n_\epsilon(p)) > 1 - \epsilon$, for $n$ large enough.

3. $|A^n_\epsilon(p)| \leq 2^{n[H(p)+\epsilon]}$ for every $n$.

4. $|A^n_\epsilon(p)| \geq (1 - \epsilon)2^{n[H(p)-\epsilon]}$, for $n$ large enough.
Proof. [1] The proof of the first claim of the theorem amounts to the following trivial chain of implications:

\[
\begin{align*}
x \in A^n(p) & \iff \frac{1}{2^n[H(p) + \epsilon]} \leq p(x_1, x_n) \leq \frac{1}{2^n[H(p) - \epsilon]} \\
& \iff -n[H(p) + \epsilon] \leq \log p(x) \leq -n[H(p) - \epsilon] \\
& \iff H(p) - \epsilon \leq -\frac{1}{n} \log p(x) \leq H(p) + \epsilon \\
& \iff \left| -\frac{1}{n} \log p(x) - H(p) \right| > \epsilon
\end{align*}
\]

where in the first step I have used the first claim of the theorem and in the second step I have used the AEP. By setting \( \delta = \epsilon \), we obtain the second claim of the theorem. [3] The proof of the third claim of the theorem amounts to the following chain of inequalities:

\[
1 = \sum_{x \in A^n} p(x) 
\geq \sum_{x \in A^n(p)} p(x) 
\geq \sum_{x \in A^n(p)} \frac{1}{2^n[H(p) + \epsilon]} 
= \left| A^n(p) \right| \frac{1}{2^n[H(p) + \epsilon]}
\]

where in the third step I have used the first claim of the theorem, namely the fact that \( p(x) \geq \frac{1}{2^n[H(p) + \epsilon]} \) for every \( x \in A^n(p) \). [4] The proof of the fourth claim of the theorem amounts to the following chain of inequalities:

\[
1 - \epsilon \leq \mathbf{P}_p(A^n(p)) \quad \text{for } n \text{ large enough} 
\leq \sum_{x \in A^n(p)} \frac{1}{2^n[H(p) - \epsilon]} 
= \left| A^n(p) \right| \frac{1}{2^n[H(p) - \epsilon]}
\]

where in the first step I have used the second claim of the theorem and in the second step I have used the first claim, namely the fact that \( p(x) \leq \frac{1}{2^n[H(p) - \epsilon]} \) for every \( x \in A^n(p) \). \( \square \)

3 Codes

Let \( \mathcal{V} \) be a finite set, whose elements are called symbols; a word on \( \mathcal{V} \) is any finite concatenation of symbols of \( \mathcal{V} \); the set of all words is denoted by \( \mathcal{V}^* \); the number of symbols concatenated in a word \( \omega \in \mathcal{V}^* \) is called the length of \( \omega \) and denoted by \( \ell(\omega) \); for any two words \( \omega_1, \omega_2 \in \mathcal{V}^* \), we say that \( \omega_2 \) is a prefix of \( \omega_1 \) iff there exists \( \omega_3 \in \mathcal{V}^* \) such that \( \omega_1 = \omega_2 \omega_3 \), i.e. \( \omega_1 \) is the concatenation of \( \omega_2 \) followed by \( \omega_3 \); we will usually assume \( \mathcal{V} = \{0, 1\} \). With this little background, we can now state the following crucial definition.

Definition 2. Consider a discrete random variable \( X \) with finite range \( \mathcal{X} \) and probability distribution \( p \). A code for \( X \) by means of an alphabet \( \mathcal{V} \) is a function \( C \) of the following form:

\[
C : \mathcal{X} \rightarrow \mathcal{V}^*
\]  

(6)
For each \( x \in \mathcal{X} \), the string \( C(x) \) is called the codeword corresponding to \( x \) with respect to the code \( C \) and the length of the word \( C(x) \) is denoted by \( \ell(C(x)) \) (or often just by \( \ell(x) \), when no confusion arises). A code is called non-singular if it is an injective function, namely the following holds for every \( x, x' \in \mathcal{X} \): if \( xx' \), then \( C(x)C(x') \). A code is called binary if \( \mathcal{V} = \{0, 1\} \); we will usually consider binary codes. The quantity \( \ell(C) \) defined as follows:

\[
\ell(C) \triangleq \mathbb{E}_p[\ell_C(X)] = \sum_{x \in \mathcal{X}} p(x)\ell_C(x)
\]

is called the expected length of the code \( C \).

4 Compression via typical sets

**Theorem 3.** Consider a sequence of i.i.d. random variables \( X_1, X_n \), with common finite range \( \mathcal{X} \). For any \( \epsilon > 0 \) and any \( n \in \mathbb{N} \) large enough, there exists a non-singular binary code \( C_\epsilon : \mathcal{X}^n \to \{0, 1\}^* \) such that its expected length is:

\[
L(C_\epsilon) = n(H(p) + \epsilon')
\]

where \( \epsilon' \) depends on \( \epsilon, n \) and the cardinality of \( \mathcal{X} \).

**Proof.** Let \( p \) be the common distribution of \( X_1, X_n \). Let \( A_\epsilon^n(p) \) be the typical set for \( p \) corresponding to \( \epsilon \) and \( n \), henceforth denoted just by \( A_\epsilon^n \). Consider an arbitrary bijection \( \alpha : A_\epsilon^n \to \{1, \ldots, |A_\epsilon^n|\} \), which assigns to each element \( x \in A_\epsilon^n \) an integer \( \alpha(x) \) between 1 and the cardinality of the set \( A_\epsilon^n \). Consider another arbitrary bijection \( \beta : \mathcal{X}^n \to \{1, \ldots, |\mathcal{X}|^n\} \) which assigns to each element of \( x \in \mathcal{X}^n \) an integer \( \beta(x) \) between 1 and the cardinality of the set \( \mathcal{X}^n \).

For each \( x \in \mathcal{X}^n \), define \( C_\epsilon(x) \) as follows: if \( x \in A_\epsilon^n \), then \( C_\epsilon(x) \leftarrow 0\omega \) (the concatenation of 0 with \( \omega \)) where \( \omega \) is the binary representation of the integer \( \alpha(x) \); if \( x \notin A_\epsilon^n \), then \( C_\epsilon(x) = 1\omega \) (the concatenation of 1 with \( \omega \)) where \( \omega \) is the binary representation of the integer \( \beta(x) \). The code \( C_\epsilon \) is trivially non-singular. Note that for every \( x \notin A_\epsilon^n \), the length \( \ell_{C_\epsilon}(x) \) can be bound as follows, where in the fourth step I have recalled that \( \omega \) is the binary representation of the integer \( \beta(x) \) which is smaller than \( |\mathcal{X}|^n \).

\[
\ell_{C_\epsilon}(x) = \ell(C_\epsilon(x)) \\
= \ell(0\omega) \\
= 1 + \ell(\omega) \\
\leq 1 + \lceil n \log |\mathcal{X}| \rceil \\
\leq 2 + n \log |\mathcal{X}|
\]

Furthermore, for every \( x \in A_\epsilon^n \), the length \( \ell_{C_\epsilon}(x) \) can be bound as follows, where in the fourth step I have recalled that \( \omega \) is the binary representation of the integer \( \alpha(x) \) which is smaller than the cardinality of \( A_\epsilon^n \) which is in turn smaller than \( 2^{n(H(p)+\epsilon)} \), as proved above.

\[
\ell_{C_\epsilon}(x) = \ell(C_\epsilon(x)) \\
= \ell(1\omega) \\
= 1 + \ell(\omega) \\
\leq 1 + \lceil n(H(p) + \epsilon) \rceil \\
\leq 2 + n(H(p) + \epsilon)
\]
I can now bound the expected length of the code $C_e$ as follows:

$$L(C_e) = \sum_{x \in \mathcal{X}^n} p(x) \ell_{C_e}(x)$$

$$= \sum_{x \in A^n} p(x) \ell_{C_e}(x) + \sum_{x \not\in A^n} p(x) \ell_{C_e}(x)$$

$$\leq (a) \sum_{x \in A^n} p(x) \left(2 + n(H(p) + \epsilon)\right) + \sum_{x \not\in A^n} p(x) \left(2 + n \log |X|\right)$$

$$= \mathbf{P}(A^n) \left(2 + n(H(p) + \epsilon)\right) + \mathbf{P}((A^n)^c) \left(2 + n \log |X|\right)$$

$$\leq (b) \left(2 + n(H(p) + \epsilon)\right) + \epsilon \left(2 + n \log |X|\right)$$

$$= n(H(p) + \epsilon) + \epsilon'$$

where in step (a) I have used both (9) and (10) and in step (b) I have used the trivial fact that $\mathbf{P}(A^n) \leq 1$ together with the fact that $\mathbf{P}((A^n)^c) \leq \epsilon$ for $n$ large enough, given that $\mathbf{P}(A^n) \geq 1 - \epsilon$, as proven above. 

\[\square\]

5 Instantaneous codes and Kraft Inequality

**Definition 3.** Let $C$ be a code for a random variable $X$ with range $\mathcal{X}$ by means of an alphabet $\mathcal{V}$. The **extension** of $C$ is the function $C^*$ defined as follows:

$$C^*: \quad \mathcal{X}^* \to \mathcal{V}^*$$

$$x_1 x_n \mapsto C^*(x_1 x_n) = C(x_1) C(x_n)$$

namely the function which maps any finite-length string $x_1 x_n$ of symbols of $\mathcal{X}$ into the string $C(x_1) C(x_n)$ obtained by concatenating in the same order the corresponding codewords. A code $C$ is called **uniquely decidable** if its extension $C^*$ is an injective function. The code $C$ is called a **prefix** or **instantaneous** or **self-punctuating** code if there are no two $x_1, x_2 \in \mathcal{X}$ such that $C(x_1)$ is a prefix of $C(x_2)$.

**Observation 1.** For concreteness, let $\mathcal{V} = \{0, 1\}$ and consider the binary tree with infinite height defined as follows:

1. each node has exactly 2 children;
2. the root is labeled with the empty string $\epsilon$;
3. each non-root node is labeled with a word $\omega \in \mathcal{V}^*$;
4. if a node is labeled $\omega$, then its left child is labeled $\omega 0$ (namely, the concatenation of $\omega$ with 0) and its right child is labeled $\omega 1$ (namely, the concatenation of $\omega$ with 1).

This infinite binary tree will be called the tree **associated** to $\{0, 1\}^*$. The extension of this construction to arbitrary alphabets $\mathcal{V}$ is of course trivial. Note that the length of a string $\omega \in \mathcal{V}^*$ is the height of the corresponding node in the tree associated with $\mathcal{V}^*$, namely the length of the path from the root to the node labeled with $\omega$. Furthermore, a word $\omega$ is a prefix of another word $\omega'$ if $\omega$ dominates $\omega'$ in the tree associated with $\mathcal{V}^*$. Thus, a code $C: \mathcal{X} \to \{0, 1\}^*$ is instantaneous if there are no two $x_1, x_2 \in \mathcal{X}$ such that $C(x_1)$ dominates $C(x_2)$ in the tree associated with $\mathcal{V}^*$.

**Theorem 4.** Kraft’s Inequality. [1] Consider a discrete random variable $X$ with finite range $\mathcal{X}$. If $C: \mathcal{X} \to \mathcal{V}^*$ is an instantaneous code for $X$ over an alphabet $\mathcal{V}$, then the following inequality holds:

$$\sum_{x \in \mathcal{X}} \left(\frac{1}{|\mathcal{V}|}\right)^{\ell_{C}(x)} \leq 1$$

(12)
Conversely, given $|X|$ integers $l_x$ with $x \in X$ such that the above inequality holds, namely:

$$\sum_{x \in X} \left(\frac{1}{|V|}\right)^{l_x} \leq 1$$

then there is an instantaneous code $C : X \to V^*$ such that $\ell_C(x) = l_x$.

**Proof.** [1] For concreteness, consider the case $V = \{0, 1\}$; the extension of the proof to the arbitrary case is trivial. Let $l = \max\{\ell_C(x) | x \in X\}$. Consider the tree associated with $V^*$, as defined in the preceding Observation. For each node $x$, let $D(x)$ be the set of nodes in the tree which are descendants of $x$ and have height $l$. Note that

$$|D(x)| = 2^{l - \ell_C(x)} \quad (13)$$

Note furthermore that the following holds, given that the code $C$ is instantaneous:

$$D(x) \cap D(x') = \emptyset \quad \text{for every } x, x' \in X \quad (14)$$

Hence:

$$2^l \geq \left| \bigcup_{x \in X} D(x) \right| = \sum_{x \in X} |D(x)| = \sum_{x \in X} 2^{l - \ell_C(x)}$$

where: in the first step, I have noted that there are $2^l$ nodes of height $l$ and I have recalled that each $D(x)$ is by definition a set of nodes of height $l$; in the second step, I have used (14) and in the third step I have sued (13). The claim follows by dividing both sides by $2^l$.

[2] Pick a node $\omega_1$ of height $l_1$ in the tree associated with $\{0, 1\}^*$, let $C(x_1) = \omega_1$ and remove all the descendants of $\omega_1$ from the tree; pick a node $\omega_2$ of height $l_2$ in the tree associated with $\{0, 1\}^*$, let $C(x_2) = \omega_2$ and remove all the descendants of $\omega_2$ from the tree; and so on. In this way, we build a prefix code with the assigned code lengths. □

### 6 Compression via Kraft’s Inequality

**Theorem 5.** Consider a discrete random variable $X$ with finite range $X$. The minimum expected length of an instantaneous code for $X$ is $H(p)$.

**Proof.** Assume that $X = \{1, \ldots, M\}$. By virtue of Kraft’s Inequality, the instantaneous code which achieves the minimum expected length is the code whose codeword lengths $l_1, \ldots, l_M$ solve the following constrained optimization problem:

- minimize: $\sum_{i=1}^{M} p_i l_i$
- subject to: $\sum_{i=1}^{M} 2^{-l_i} \leq 1$
- $l_i \geq 0$ for $i = 1, \ldots, M$

We can solve this problem by means of the method of Lagrange multipliers. The corresponding Lagrangian is as follows:

$$\Lambda(l, \lambda) = \sum_{i=1}^{M} p_i l_i + \lambda \left( \sum_{i=1}^{M} 2^{-l_i} - 1 \right) \quad (15)$$

and its derivatives are as follows:

$$\frac{\partial \Lambda(l, \lambda)}{\partial l_i} = p_i - \lambda 2^{-l_i} \ln 2 \quad (16)$$
Thus, the condition \( \frac{\partial \Lambda(l, \lambda)}{\partial l_i} = 0 \) holds iff the following holds:

\[
2^{-l_i} = \frac{p_i}{\lambda \ln 2}
\] (17)

By imposing that the constrains are satisfied, I get the following:

\[
1 = \sum_{i=1}^{M} 2^{-l_i} = \sum_{i=1}^{M} \frac{p_i}{\lambda \ln 2} = \frac{1}{\lambda \ln 2}
\] (18)

from which I derive that \( \lambda = \frac{1}{\ln 2} \). By replacing this expression for \( \lambda \) in (17), I conclude that \( p_i = 2^{-l_i} \) and thus:

\[
l_i = -\log p_i = \log \frac{1}{p_i}
\] (19)

Thus, the optimal code \( C_{opt} \) has code lengths \( l_i = \log \frac{1}{p_i} \) and its expected length is:

\[
\ell(C_{opt}) = \sum_{i=1}^{M} p_i l_i = \sum_{i=1}^{M} p_i \log \frac{1}{p_i} = H(p)
\] (20)

namely, the entropy of \( X \). \( \square \)