Contents

Chapter 1

Introduction

1.1 Entropy

Let X be a random variable with probability mass function $p(x)$, then the **entropy** of X is defined as

$$H(X) = \mathbb{E}[-\log(p(X))] = -\sum_{x \in \mathcal{X}} p(x) \log(p(x))$$

which intuitively measures the uncertainty of a single variable. Depending one the base of the logarithm, the entropy is measured in bits, for base 2, nats, for base e. Entropy can also be viewed as the average amount information revealed after sampling X. We can define conditional entropy of X given that $Y = y$ to be

$$H(X|Y = y) = -\sum_{x \in \mathcal{X}} p_{X|Y}(x|y) \lg\left(\frac{p_{XY}(x, y)}{p_Y(y)}\right)$$

and conditional entropy of X given Y is

$$\begin{aligned} H(X|Y) &= \sum_{y \in \mathcal{Y}} p_Y(y) H(X|Y = y) \\ &= -\sum_{y} \sum_{x} p_{XY}(x, y) \lg\left(\frac{p_{XY}(x, y)}{p_Y(y)}\right) \end{aligned}$$

Lastly, the joint entropy to variables is defineds

$$H(X, Y) = \mathbb{E}_{X,Y}[-\log(p_{XY}(X, Y))] = -\sum_{x,y} p_{XY}(x, y) \lg(p_{XY}(x, y))$$

From now on we omit the subscript for the PMFs unless it can not be inferred from the context.

Proposition 1.1 (Chain rule for entropy). *For any two random variables X and Y*

$$H(X, Y) = H(X) + H(Y|X)$$

furthermore if Z is another random variable then

$$H(X, Y|Z) = H(X|Z) + H(Y|X, Z)$$

which then can be used to generalize the chain rule

$$H(X_1, \dots, X_n) = \sum i = 1^n H(X_i|X_{i-1}, \dots, H(X_1))$$

Proof. For the conditional case

$$
\begin{aligned}
H(X|Z) &= -\sum_{x,z} p(x,z) \lg\left(\frac{p(x,z)}{p(z)}\right) \\
H(Y|X,Z) &= -\sum_{x,y,z} p(x,y,z) \lg\left(\frac{p(x,y,z)}{p(x,z)}\right) \\
\implies H(X|Z) + H(Y|X,Z) &= -\sum_{x,y,z} p(x,y,z) \lg\left(\frac{p(x,y,z)}{p(z)}\right) \\
&= H(X,Y|Z)
\end{aligned}
$$

1.2 Mutual information

Mutual information is the reduction in entropy due to another random variable.

$$
\begin{aligned}
I(X;Y) &= H(X) - H(X|Y) \\
&= \mathbb{E}_{x,y}\left[\lg\left(\frac{p(X,Y)}{p(X)p(Y)}\right)\right] \\
&= \sum_{x}\sum_{y} p(x,y) \lg\left(\frac{p(x,y)}{p(x)p(y)}\right) \\
&= H(Y) - H(Y|X) = I(Y;X)
\end{aligned}
$$

Proposition 1.2. *$I(X;Y)$ is zero if and only if X and Y are independent.*

For conditional mutual information we have

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

Proposition 1.3 (Chain rule for mutual information). *For a random variable Y and random variables $X_1, \ldots, X_n$ we have*

$$
I(X_1, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y | X_{i-1}, \ldots, X_1)
$$

Proof. We have

$$
\begin{aligned}
I(X_1, \ldots, X_n; Y) &= H(X_1, \ldots, X_n) - H(X_1, \ldots, X_n | Y) \\
&= \sum_{i=1}^{n} H(X_i | X_{i-1}, \ldots, X_1) - H(X_i | X_{i-1}, \ldots, X_1, Y) \\
&= \sum_{i=1}^{n} I(X_i; Y | X_{i-1}, \ldots, X_1)
\end{aligned}
$$

1.3 Channel Capacity

A *communication channel* is a system in which output depends probabilistically on its input. It is characterized by a probability transition matrix $p(y|x)$. **Capacity** of a communication channel with input X and output Y is defined as

$$C = \max_{p(x)} I(X;Y)$$

1.4 Relative entropy

Relative entropy or *Kullback–Leibler divergence* measures how one probability distribution differs from another.

$$D(p||q) = \mathbb{E}_{p(x)}\left[\lg\left(\frac{p(X)}{q(X)}\right)\right] = \sum_x p(x)\lg\left(\frac{p(x)}{q(x)}\right)$$

Even though it is not a metric, if $D(p||q) = 0 \implies p = q$.

Note that

$$I(X;Y) = \sum_{x,y} p(x,y)\lg\left(\frac{p(x,y)}{p(x)p(y)}\right) = D(p(x,y)||p(x)p(y))$$

Conditional relative entropy is defined as

$$\begin{aligned} D(p(y|x)||q(y|x)) &= \mathbb{E}_{p(x,y)}\left[\lg\left(\frac{p(Y|X)}{q(Y|X)}\right)\right] \\ &= \sum_x p(x)\sum_y p(y|x)\lg\left(\frac{p(y|x)}{q(y|x)}\right) \end{aligned}$$

Similarly we define the following chain rule

$$D(p(x,y)||q(x,y)) = D(p(x)||q(x)) + D(p(y|x)||q(y|x))$$

1.5 Convex function and inequalities

A function f is said to be convex over an interval $[a,b]$ if for every $x_1, x_2 \in \,]a,b[$ and $0 \leq \lambda \leq 1$

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda)f(x_2)$$

f is said to be strictly convex if equality holds only if $\lambda = 0, 1$.

Theorem 1.4. *If f is twice differentiable and has non-negative (positive) second derivative over an interval, then f is convex (strictly convex) over that interval.*

Theorem 1.5 (Jensen's inequality). *If f is a convex function and X is a random variable*

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$$

Moreover, if f is strictly convex, the equality implies that $X = \mathbb{E}[X]$ with probability 1.

Corollary 1.6. *The followings can be shown using the Jensen's inequality*

1. *For non-negative numbers $a_1, \dots, a_n$ and $b_1, \dots, b_n$*

$$\sum a_i \log\left(\frac{a_i}{b_i}\right) \geq \left(\sum a_i\right) \log\left(\frac{\sum b_i}{\sum a_i}\right)$$

equality holds if and only $\frac{a_i}{b_i} = c, \quad \forall i$. This is called log sum inequality.

2. *$D(p||q) \geq 0$ and equality holds when $p = q$.*

3. *$I(X;Y) \geq 0$ and equality holds when X and Y are independent.*

4. *$D(p(y|x)||q(y|x)) \geq 0$ and equality holds when $p(y|x) = q(y|x)$ for all x and y such that $p(x) > 0$.*

5. *$I(X;Y|Z) \geq 0$ and equality holds when X and Y are conditionally independent given Z.*

Proof. 1. Suppose $\lambda_i = b_i$, $x_i = \frac{a_i}{b_i}$, and $f(x) = x \log x$ then

$$\begin{aligned} \frac{\sum \lambda_i f(x_i)}{\sum \lambda_i} &= \frac{\sum a_i \log \frac{a_i}{b_i}}{\sum b_i} \\ &\geq \frac{\sum a_i}{\sum b_i} \log\left(\frac{\sum a_i}{\sum b_i}\right) \\ \implies \sum a_i \log\left(\frac{a_i}{b_i}\right) &\geq \left(\sum a_i\right) \log\left(\frac{\sum b_i}{\sum a_i}\right) \end{aligned}$$

For the Jensen inequality, equality holds when $x_1 = \cdots = x_n$ and thus $\frac{a_i}{b_i} = c, \forall i$.

2. Using the log sum inequality for $a_i = p(x_i)$ and $b_i = q(x_i)$

$$\begin{aligned} \sum_i p(x_i) \log\left(\frac{p(x_i)}{q(x_i)}\right) &\geq \left(\sum p(x_i)\right) \log\left(\frac{\sum p(x_i)}{\sum q(x_i)}\right) \\ &= 0 \end{aligned}$$

equality holds when $p(x) = cq(x)$, and since both are PMFs $c = 1$.

3. we know that

$$I(X;Y) = D(p(x,y)||p(x)p(y)) \geq 0$$

and equality holds when $p(x,y) = p(x)p(y)$ which means X and Y are independent.

4. Using the log sum inequality for $a_i = p(y_i|x)$ and $b_i = q(y_i|x)$

$$\begin{aligned} \sum_x p(x) \sum_{y_i} p(y_i|x) \log\left(\frac{p(y_i|x)}{q(y_i|x)}\right) &\geq \sum_x p(x) \left(\sum p(y|x_i)\right) \log\left(\frac{\sum p(y|x_i)}{\sum q(y_i|x)}\right) \\ &= 0 \end{aligned}$$

equality holds when $p(y|x) = q(y|x)$ for all y and x with $p(x) > 0$.

5. Since

$$I(X;Y|Z) = D(p(x,y|z)||p(x|z)p(y|z)) \geq 0$$

and equality holds when X and Y are independent given Z. ∎

Theorem 1.7. *For any random variable X*

$$H(X) \leq \log|X|$$

with equality if and only if X has a uniform distribution.

Proof. Let $u(X)$ be the uniform distribution on X. Then

$$\begin{aligned} D(p||u) &= \sum p(x) \log\left(\frac{p(x)}{u(x)}\right) \\ &= \sum p(x) \log(p(x)) + \log(|X|) \\ &= -H(X) + \log|X| \geq 0 \\ \implies \log|X| &\geq H(X) \end{aligned}$$

Theorem 1.8 (Conditioning reduces entropy).

$$H(X|Y) \geq H(X)$$

However this is on average. That is, $H(X|Y=y)$ might be greater than $H(X)$.

Proof. Mutual information $I(X;Y)$ is greater than zero. ∎

Corollary 1.9 (Independence bound on entropy).

$$H(X_1, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i)$$

Theorem 1.10 (Convexity of relative entroy). *For any two pairs probability mass functions (p_1, q_1) and (p_2, q_2)*

$$D(\lambda p_1 + (1-\lambda)p_1||\lambda q_1 + (1-\lambda)q_1) \leq \lambda D(p_1||q_1) + (1-\lambda)D(p_2||q_2)$$

for all $0 \leq \lambda \leq 1$.

Proof. Note that using the log sum inequality on each term

$$(\lambda p_1 + (1-\lambda)p_2) \log\left(\frac{\lambda p_1 + (1-\lambda)p_2}{\lambda q_1 + (1-\lambda)q_2}\right) \leq \lambda p_1 \log\frac{p_1}{q_1} + (1-\lambda)\log\frac{p_2}{q_2}$$

Theorem 1.11 (Concavity of entropy). *$H(X)$ is a concave function of its distribution, $p(x)$.*

Proof.

$$H(X) = \log|X| - D(p||u)$$

Theorem 1.12. *The mutual information $I(X;Y)$ is a concave function of $p(x)$ for fixed $p(y|x)$ and a convex function of $p(y|x)$ for fixed $p(x)$*

Definition (Markov chain): Let X, Y, Z be random variables are said to form a Markov chain in that order $X \to Y \to Z$ if the conditional distribution of Z depends only Y and is conditionally independent of X. Specifically

$$p(x, y, z) = p(x)p(y|x)p(z|y)$$

For example $Z = f(Y)$ then $X \to Y \to Z$ is a Markov chain. Note that

$$X \to Y \to Z \implies Z \to Y \to Z$$

and hence we can write $X \leftrightarrow Y \leftrightarrow Z$.

Theorem 1.13 (Data processing inequality). *If $X \to Y \to Z$ is a Markov chain, then*

$$I(X;Y) \geq I(X;Z)$$

equality happens if $I(X;Y|Z) = 0$ which implies $X \to Z \to Y$.

1.6 Sufficient statistics

Let $\{f_\theta(x)\}_\theta$ be a family of PMSs and let X be a sample from a distribution in this family. Let $T(X)$ be any statistics. Then, $\theta \to X \to T(X)$ is Markov chain and hence

$$I(\theta; X) \geq I(\theta; T(X))$$

$T(X)$ is sufficient statistics for parameter θ if the conditional distribution of X given $T(X)$ does not depend on θ. Therefore, for a sufficient statistics $\theta \to T(X) \to X$ and thus the data processing inequality becomes an equality

$$I(\theta; X) = I(\theta; T(X))$$

A statistics $T(X)$ is a minimal sufficient statistics relative to $\{f_\theta(x)\}$ if it is a function of every other sufficient statistics U. Equivalently

$$\theta \to T(X) \to U(X) \to X$$

We observe a random variable Y and we guess the correlated variable X using $\hat{X} = f(Y)$ for some function f. Then we wish to know the probability of error

$$P_e = \mathbb{P}\left(X \neq \hat{X}\right)$$

Fano's inequality gives bound on P_e.

Theorem 1.14 (Fano's inequality). *For any estimator $\hat{X}$ such that $X \to Y \to \hat{X}$ with $P_e = \mathbb{P}\left(X \neq \hat{X}\right)$ we have*

$$H(P_e) + P_e \lg|X| \geq H\left(X|\hat{X}\right) \geq H(X|Y)$$

and thus

$$\begin{aligned} 1 + P_e \lg|X| &\geq H(X|Y) \\ \implies P_e &\geq \frac{H(X|Y) - 1}{\lg|X|} \end{aligned}$$

Intuitively, this inequality says that if Y does not give much information about X then P_e is greater than when Y has a lot information about X.

Proof. Let E be the random variable with

$$E = \begin{cases} 1 & X = \hat{X} \\ 0 & X \neq \hat{X} \end{cases}$$

then

$$\begin{aligned} H\left(E, X|\hat{X}\right) &= H\left(E|\hat{X}\right) + H\left(X|E, \hat{X}\right) \\ &= H\left(X|\hat{X}\right) + H\left(E|X, \hat{X}\right) = H\left(X|\hat{X}\right) \end{aligned}$$

therefore

$$\begin{aligned} H\left(X|\hat{X}\right) &= H\left(E|\hat{X}\right) + H\left(X|E, \hat{X}\right) \\ &\leq H(E) + H\left(X|E = 0, \hat{X}\right)\mathbb{P}(E = 0) + H\left(X|E = 1, \hat{X}\right)\mathbb{P}(E = 1) \\ &\leq H(P_e) + P_e H(X) \\ &\leq H(P_e) + P_e \lg|X| \end{aligned}$$

and by data processing inequality

$$H\left(X|\hat{X}\right) \geq H(X|Y)$$

Corollary 1.15. *For any two random variables X, Y let $p = \mathbb{P}(X \neq Y)$ then*

$$H(p) + p \lg|X| \geq H(X|Y)$$

Proof. Let $\hat{X} = Y$ in Fano's inequality. ∎

Corollary 1.16. *Let $P_e = \mathbb{P}\left(X \neq \hat{X}\right)$ where $\hat{X} : \mathcal{Y} \to \mathcal{X}$ then*

$$H(P_e) + P_e \lg(|X| - 1) \geq H(X|Y)$$

Lemma 1.17. *If X, X' are i.i.d with entropy $H(X)$ then*

$$\mathbb{P}(X = X') \geq 2^{-H(X)}$$

Proof.

$$\mathbb{P}(X = X') = \sum_x p^2(x) = \sum_x p(x) 2^{\lg p(x)} \geq 2^{\sum p(x) \lg p(x)} = 2^{-H(X)}$$

Corollary 1.18. *Let X, X' be independent variables with $X \sim p(x)$ and $X' \sim r(x)$, $x, x' \in \mathcal{X}$ then*

$$\begin{aligned} \mathbb{P}(X = X') &\geq 2^{-H(p) - D(p||r)} \\ &\geq 2^{-H(r) - D(r||p)} \end{aligned}$$

Example 1.1. We will prove the fact there are infinitely many primes. Let

$$\pi(x) = |\{p \leq x \,|\, p \text{ is a prime}\}|$$

Let $N \sim \text{Unif}\{1, \ldots, n\}$ then by the prime factorization theorem

$$N = \prod_{i=1}^{\pi(n)} p_i^{X_i}$$

where X_i are random variables.

$$2^{X_i} \leq p_i^{X_i} \leq N \leq n \implies X_i \leq \lg n$$

Furthemore

$$H(N) = H\big(X_1, \ldots, X_{\pi(n)}\big) \leq \sum_{i=1}^{\pi(n)} H(X_i)$$

therefore

$$\lg n \leq \sum_{i=1}^{\pi(n)} H(X_i) \leq \pi(n) \lg(\lg n + 1)$$

implying that

$$\pi(n) \geq \frac{\lg n}{\lg(\lg n + 1)}$$

hence $\pi(n) \to \infty$ as $n \to \infty$.

1.7 Shearer inequality

Let $(X_1, \ldots, X_n)$ be a random vector and let $A_1, \ldots, A_k \subset \mathbb{N}_n$ be such that every integer $i \in \mathbb{N}_n$ lies in at least r of them . Then

$$H(X_1, \ldots, X_n) \geq \frac{1}{r} \sum_{i=1}^{n} H((X_j)_{j \in A_i})$$

Example 1.2. Let S be a set of distincts points in $\mathbb{R}^3$.

- n_1 be the number distinct point after projection onto $x = 0$ plane.
- n_2 be the number distinct point after projection onto $y = 0$ plane.
- n_3 be the number distinct point after projection onto $z = 0$ plane.

Then,

$$n^2 \leq n_1 n_2 n_3$$

Example 1.3. Let $G(V, E)$ be an undirected graph and let t be the number of triangles in G. Then

$$t \leq \frac{1}{6}(2l)^{\frac{3}{2}}$$

where $l = |E|$.

Chapter 2

Source Coding

We have an information source modeled by a random variable $\mathcal{X} \ni X \sim p_X$. A source code C is a function from $\mathcal{X} \to D^*$ where D^* us the set of all finite length strings over the alphabet D.

Example 2.1. Let $\mathcal{X} = \{\text{red}, \text{blue}\}$ and $D = \{0, 1\}$. A possible source code might be

$$C(\text{red}) = 00 \quad C(\text{blue}) = 110$$

Definition: Length of the codeword of $x \in \mathcal{X}$ is

$$l(x) = |C(x)|$$

and the average length of code C

$$L(C) = \mathbb{E}[l(X)] = \sum_{x \in \mathcal{X}} p(x) l(x)$$

Definition: C is non-singular if

$$C(x_1) = C(x_2) \implies x_1 = x_2$$

We want a source code C to be invertible with minimum average length. Furthermore, we usually want to code a sequence $x_1, \ldots, x_n$. One way to this is to introduce a character ‘,’ that is not in D then

$$C(x_1, \ldots, x_n) = C(x_1), C(x_2), \ldots, C(x_n)$$

Another way is to use instantaneous codes – also called prefix-free codes.

Definition: Extension of C is

$$C^*(x_1, \ldots, x_n) = C(x_1) \ldots C(x_n)$$

Then C is uniquely decodable if for each pair of different sequences $x_1, \ldots, x_n \neq y_1, \ldots, y_m$

$$C^*(x_1, \ldots, x_n) \neq C^*(y_1, \ldots, y_m)$$

Equivalently, C^* is non-singular.

Note that, decoding a uniquely decodable code might not be possible until the very end of stream, which is problematic. To do away with we introduce the prefix-free codes.

Definition: A prefix-free code C is such that no codeword is a prefix of another codeword,

Example 2.2. Let $\mathcal{X} = \{1, 2, 3, 4\}$ and $D = \{0, 1\}$. Consider

$$C_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad C_2 = \begin{bmatrix} 0 \\ 010 \\ 01 \\ 10 \end{bmatrix} \quad C_3 = \begin{bmatrix} 10 \\ 00 \\ 11 \\ 110 \end{bmatrix} \quad C_4 = \begin{bmatrix} 0 \\ 10 \\ 110 \\ 111 \end{bmatrix}$$

then C_1 is a singular code, C_2 is non-singular but it is not uniquely decodable, C_3 is non-singular and uniquely decodable but it is not prefix-free, lastly, C_4 is a non-singular prefix-free code.

2.1 Kraft inequality

Theorem 2.1. *For any instantaneous code over alphabet D with size d that has codeword length $l_1, l_2, \ldots, l_m$*

$$\sum_{i=1}^{m} d^{-l_i} \leq 1$$

Conversely, if $l_1, \ldots, l_m$ satisfy the equation above, then there exists an instantaneous code with those codeword length.

Proof.

Remark 1. If m is inifinte but countable the Kraft inequality and its converse still hold.

2.2 Minimizing the average length

$$\begin{aligned}
\bar{L}_{opt} &= \min_{C \in PF} L(C) \\
&= \min_{C \in PF} \sum_x p(x) l(x) \\
&= \min_{l(x)} \sum_x p(x) l(x) && \text{subject to } \begin{cases} l : \mathcal{X} \to \mathbb{N} \\ \sum d^{-l(x)} \leq 1 \end{cases} \\
&= \min_{q(x)} - \sum_x p(x) \log_d(q(x)) && \text{subject to } \begin{cases} q : \mathcal{X} \to \{\frac{1}{d}, \frac{1}{d^2}, \ldots\} \\ \sum q(x) \leq 1 \end{cases} \\
&\geq \min_{q(x)} - \sum_x p(x) \log_d(q(x)) = A && \text{subject to } \begin{cases} q(x) \geq 0 \\ \sum q(x) \leq 1 \end{cases}
\end{aligned}$$

let B be

$$B = \min_{q^*(x)} - \sum_x p(x) \log_d(q^*(x)) \qquad \text{subject to } \begin{cases} q^*(x) \geq 0 \\ \sum q^*(x) \leq 1 \end{cases}$$

It is clear that $A \leq B$. We claim that $B \leq A$ and hence $A = B$. Let $S(q) = \sum q(x)$ and then let $q^*(x) = \frac{q(x)}{S(q)}$. Then

$$\begin{aligned}
A &= \min_{q(x)} -\sum_x p(x)\log_d(q(x)) && \text{subject to } \begin{cases} q(x) \geq 0 \\ \sum q(x) \leq 1 \end{cases} \\
&= \min_{q^*(x)} -\sum_x p(x)\log_d(q^* S(q)(x)) && \text{subject to } \begin{cases} q^*(x) \geq 0 \\ \sum q^*(x) = 1 \end{cases} \\
&= \min_{q^*(x)} -S(q) - \sum_x p(x)\log_d(q^*(x)) && \text{subject to } \begin{cases} q^*(x) \geq 0 \\ \sum q^*(x) = 1 \end{cases} \\
&= B - S(q) \geq B
\end{aligned}$$

Therefore,

$$\begin{aligned}
\bar{L}_{opt} &\geq B \\
&= \min_{q(x)} -\sum_x p(x)\log_d(q(x)) && \text{subject to } \begin{cases} q(x) \geq 0 \\ \sum q(x) \leq 1 \end{cases} \\
&= \min_{q(x)} \sum_x p(x)\left(\log\left(\frac{p(x)}{q(x)}\right) - \log_d(q^*(x))\right) \\
&= \min_{q(x)} D(p||q) + H_d(X)
\end{aligned}$$

Since $D(p||q) \geq 0$ therefore

$$\bar{L}_{opt} \geq H_d(X)$$

and the equality holds iff

$$l(x) = -\log(p(x)) \in \mathbb{N} \implies p(x) = \left\{\frac{1}{d}, \frac{1}{d^2}, \ldots\right\}$$

To get an upperbound for $\bar{L}_{opt}$ consider Shannon-Fano code. Shannon-Fano code assigns a codeword of length $l(x) = \lceil -\log_d p(x) \rceil$. Shannon-Fano code satisfy the Kraft's inequality

$$\begin{aligned}
\sum_{x \in \mathcal{X}} d^{-l(x)} &= \sum_{x \in \mathcal{X}} d^{-\lceil -\log_d p(x) \rceil} \\
&= \sum_{x \in \mathcal{X}} d^{\lfloor \log_d p(x) \rfloor} \leq \sum_{x \in \mathcal{X}} d^{\log_d p(x)} = 1
\end{aligned}$$

Hence, there exists a prefix-free code C_{Sh-F} with such a codeword lengths.

$$\begin{aligned}
-\log_d p(x) \leq \lceil -\log_d p(x) \rceil &< \log_d p(x) \\
H_d(X) \leq L(C_{Sh-F}) &< H_d(X) + 1 \\
\implies H_d(X) \leq \bar{L}_{opt} \leq L(C_{Sh-F}) &< H_d(X) + 1
\end{aligned}$$

In multishot coding, we encode a block input as oppose to only one sample. Let $\underline{x} \in \mathcal{X}^n$ be a block of length n then

$$H_d(\underline{X}) = H_d(X_1, \ldots, X_n) \leq \bar{L}_{opt}^{(n)} < H_d(X_1, \ldots, X_n) + 1$$

assuming that the source is i.i.d.

$$H_d(X) \leq \frac{1}{n}\bar{L}_{opt}^{(n)} < H_d(X_1, \ldots, X_n) + \frac{1}{n}$$

Therefore, as $n \to \infty$ the average length for each input symbol approaches the entopy.

Example 2.3. Let $\mathcal{X} = \{A, B, C\}$ with distribution $p(A) = p(B) = p(C) = \frac{1}{3}$ and $D = \{0, 1\}$ then

$$H(X) \simeq 1.58$$

for Shannon-Fano code, all the codewords have length 2. Hence

$$C_{Sh-F} = \{00, 01, 10\} \implies L(C_{Sh-F}) = 2$$

Huffman code which will be described later produces

$$C_{Huff} = \{0, 10, 11\} \implies L(C_{Huff}) = 1.67$$

Note that Huffman code is optimal but Shannon-Fano code is not as Multishot Huffman codes approaches the value of entropy of X.

2.3 Kraft's inequality for uniquely decodable

Theorem 2.2. *The set of code lengths of unique decodable over an alphabet D with size d satisfy*

$$\sum_{x \in \mathcal{X}} d^{-l(x)}$$

The converse is implied by the converse of Kraft's inequality for prefix-free codes.

2.4 Huffman Code

The algorithm is

1. Sort the PMF in the decreasing order.
2. combine the least two.
3. continue until you have two symbols.
4. assign 0 to the left one and assign 1 to right one.
5. Backtrack and for each two symbol combined, append 0 to the left one and append 1 to the one.

For Huffman algorithm to work for $d \geq 3$ we must have

$$m = 1 + k(d - 1)$$

we can add symbols with 0 probability.

Example 2.4. We wish to guess $x \in \mathcal{X}$ with the least number of question of form " is $x \in S$ for some subset of $\mathcal{X}$". We also know the distribution p_X.

Lemma 2.3. *Every uniquely decodable code is a sequence of questions and vice versa.*

$C_{opt}^{(m-1)}(\mathbb{P}')$		C_{ext}
$p_1 \to c'_1, l'_1$		$p_1 \to c'_1, l'_1$
$p_2 \to c'_2, l'_2$		$p_2 \to c'_2, l'_2$
$\vdots$	$\xrightarrow{extend}$	$\vdots$
$p_{m-1} + p_m \to c'_{m-1}, l'_{m-1}$		$p_{m-1} \to c'_{m-1}\vert\vert 0, l'_{m-1}+1$ $p_m \to c'_m\vert\vert 1, l'_{m-1}+1$

$C_{can}^{(m)}(\mathbb{P})$		C_{mrg}
$p_1 \to c_1, l_1$		$p_1 \to c_1, l_1$
$p_2 \to c_2, l_2$		$p_2 \to c_2, l_2$
$\vdots$	$\xrightarrow{merge}$	$\vdots$
$p_{m-1} \to c_{m-1}, l_{m-1}$ $p_m \to c_m, l_m = l_{m-1}$		$p_{m-1} + p_m \to c_m[1 : l_m - 1], l_m - 1$

2.4.1 Optimality of Huffman code

Lemma 2.4. *For an optimal code we must have*

- $p(x) \geq p(y) \implies l(x) \leq l(y)$.
- *for an instantaneous, the longest two codes have the same length.*

Furthermore, there exists an instantaneous such that the longest two codes differ in the last bit.

Proof. 1) do algebra, 2) and 3) prefix tree ∎

A code that satisfy the preceding properties is called cannonical code. Let $\mathcal{X}$ be some alphabet with size m and PMF $\mathbb{P} = (p_1, \ldots, p_m)$ with $p_1 \geq \cdots \geq p_m$. Consider the Huffman reduction algorithm

$$\mathbb{P}' = (p_1, \ldots, p_{m-2}, p_{m-1} + p_m)$$

definde over $\mathcal{X}'$ with $|\mathcal{X}'| = m - 1$. Suppose, $C_{can}^{(m)}(\mathbb{P})$ is a cannonical code over $\mathcal{X}$ and $C_{opt}^{(m-1)}(\mathbb{P}')$ be an optimal code over $\mathcal{X}'$. Therefore,

$$L(C_{ext}) = L\Big(C_{opt}^{(m-1)}\Big) + (p_{m-1} + p_m)$$

similarly for merging Thus,

$$L(C_{mrg}) = L\big(C_{can}^{(m)}\big) - (p_{m-1} + p_m)$$

hence

$$\begin{aligned} L(C_{mrg}) + L(C_{ext}) &= L\Big(C_{opt}^{(m-1)}\Big) + L\big(C_{can}^{(m)}\big) \\ \implies \Big(L(C_{ext}) - L\Big(C_{opt}^{(m-1)}\Big)\Big) &+ \big(L(C_{mrg}) - L\big(C_{can}^{(m)}\big)\big) = 0 \end{aligned}$$

but both terms are greater than zero(by Optimality and cannonical). Therefore, $L(C_{ext}) = L\Big(C_{opt}^{(m-1)}\Big)$ and $L(C_{mrg}) = L\Big(C_{can}^{(m)}\Big)$. Huffman code is like merging and then extending. Therefore, Huffman code is an optimal uniquely decodable code.

Chapter 3

Asymptotic Equipartition Property

3.1 AEP and the typical set

Example 3.1. Suppose $X_i \sim \text{Bernoulli}(p)$ are i.i.d. then

$$\begin{aligned}
\mathbb{P}(X_1 + \cdots + X_n = k) &= \binom{n}{k} p^k (1-p)^{n-k} \\
&\simeq \frac{\left(\frac{n}{e}\right)^n}{\left(\frac{k}{e}\right)^k \left(\frac{n-k}{e}\right)^{n-k}} p^k (1-p)^{n-k} \\
&= \left(\frac{n}{k}\right)^k \left(\frac{n}{n-k}\right)^{n-k} p^k (1-p)^{n-k}
\end{aligned}$$

Let $k = np$

$$= p^{-np}(1-p)^{-n(1-p)} p^{np} (1-p)^{n(1-p)}$$

Theorem 3.1 (AEP theorem). *Suppose $X_1, X_2, \cdots \sim f(X)$ are i.i.d. then*

$$-\frac{1}{n} \lg f(X_1, \ldots, X_n) \xrightarrow{\mathbb{P}, a.s.} H(X)$$

Proof. By the weak/strong law of large numbers

$$\begin{aligned}
-\frac{1}{n} \lg f(X_1, \ldots, X_n) &= -\frac{1}{n} \lg(f(X_1) \ldots f(X_n)) \\
&= -\frac{1}{n} \sum \lg f(X_i) \\
&\xrightarrow{\mathbb{P}, a.s.} -\mathbb{E}[\lg f(X)] = H(X)
\end{aligned}$$

Definition: A typical set $A_\epsilon^{(n)}$

$$A_\epsilon^{(n)} = \left\{ (x_1, \ldots, x_n) \middle| \left| -\frac{1}{n} \lg f(X_1, \ldots, X_n) \right| - H(X) < \epsilon \right\}$$

hence if $\bar{x} \in A_\epsilon^{(n)}$ then

$$2^{-n(H(X)+\epsilon)} \leq \mathbb{P}(\bar{x}) \leq 2^{-n(H(X)-\epsilon)}$$

Proposition 3.2. *1. For sufficiently large n*

$$\mathbb{P}\big(\bar{x} \in A_\epsilon^{(n)}\big) \geq 1 - \epsilon$$

2. For all n

$$\big|A_\epsilon^{(n)}\big| \leq 2^{n(H(X)+\epsilon)}$$

and for sufficiently large n

$$\big|A_\epsilon^{(n)}\big| \geq (1-\epsilon)2^{n(H(X)-\epsilon)}$$

Proof. 1. By the AEP theorem there exists N such that for all $n \geq N$

$$\mathbb{P}\left(\left|-\frac{1}{n}\lg f(X_1,\ldots,X_n) - H(X)\right| < \epsilon\right) \geq 1-\epsilon$$

which means for sufficiently large n

$$\mathbb{P}\big(\bar{x} \in A_\epsilon^{(n)}\big) \geq 1 - \epsilon$$

2. For each $\bar{x} \in A_\epsilon^{(n)}$

$$2^{-n(H(X)+\epsilon)} \leq \mathbb{P}(\bar{x})$$

therfore

$$\big|A_\epsilon^{(n)}\big|2^{-n(H(X)+\epsilon)} \leq \mathbb{P}\big(\bar{x} \in A_\epsilon^{(n)}\big) \leq 1$$

and similarly from the last result, for sufficiently large n

$$1-\epsilon \leq \mathbb{P}\big(\bar{x} \in A_\epsilon^{(n)}\big) \leq \big|A_\epsilon^{(n)}\big|2^{-n(H(X)-\epsilon)}$$

By AEP we prove the compression theorem, that is, there exists C such that $L(C) = H(X)$. This code works as follow

- Since there are at most $2^{n(H(X)+\epsilon)}$ sequences in $A_\epsilon^{(n)}$ then we can code them all with codewords of length $n(H(X)+\epsilon)+1$ bits.
- There are at most $|\mathcal{X}|^n$ sequnces not in $A_\epsilon^{(n)}$ and hence we can code them with codewords of length $n\lg|\mathcal{X}|+1$ bits.
- To be able uniquely decode typical sequences from atypical ones, prefix a 1 for typicals and prefix a 0 for atypical sequences.

The average length of the above coding scheme is

$$\begin{aligned}
L(C) &\leq \sum_{\bar{x}\in A_\epsilon^{(n)}} \mathbb{P}(\bar{x})(n(H(X)+\epsilon)+2) + \sum_{\bar{x}\notin A_\epsilon^{(n)}} \mathbb{P}(\bar{x})(n\lg|\mathcal{X}|+2) \\
&= (n(H(X)+\epsilon)+2)\big(1-\mathbb{P}\big(\bar{x}\notin A_\epsilon^{(n)}\big)\big) + (n\lg|\mathcal{X}|+2)\mathbb{P}\big(\bar{x}\notin A_\epsilon^{(n)}\big) \\
&= nH(X) + n\epsilon + 2 + n(\lg\mathcal{X} - H(X) - \epsilon)\mathbb{P}\big(\bar{x}\notin A_\epsilon^{(n)}\big)
\end{aligned}$$

Let n be sufficiently large that $\mathbb{P}\left(\bar{x}\notin A_\epsilon^{(n)}\right) \leq \epsilon$

$$\begin{aligned}
&\leq nH(X) + n\epsilon + 2 + n(\lg\mathcal{X} - H(X) - \epsilon)\epsilon \\
&= n(H(X)+\epsilon') \qquad \epsilon' = \epsilon + \frac{2}{n} + \epsilon\lg\mathcal{X} - \epsilon H(X) - \epsilon^2
\end{aligned}$$

Where ϵ' can be made arbitrarily small and thus

$$L(C) \xrightarrow[n\to\infty]{\epsilon\to 0} nH(X)$$

3.2 High-probability set

It is clear that $A_\epsilon^{(n)}$ is a fairly small set with high probability. We will argue that it has the about as small as the smallest set with high probability.

Definition: Let $B_\epsilon^{(n)}$ be the smallest set such that

$$\mathbb{P}\big(\bar{x} \in B_\epsilon^(n)\big) \leq 1 - \epsilon$$

Theorem 3.3. *Let $X_1, \cdots \sim f(X)$ be i.i.d and (condition on ϵ) then for any $\delta > 0$*

$$\frac{1}{n} \lg \big| B_\epsilon^{(n)} \big| > H(X) - \delta$$

for sufficiently large n. Therefore, $B_\epsilon^{(n)}$ contains at least $2^{nH(X)}$ elements hence A_ϵ^n has about the same number elements as the high-probability set.

3.3 Stochastic Processes

For the general case when X_i are not independent, we can assume they form an stochastic process. A stationary stochastic process is invariant to time shifts

$$f(x_1, \ldots, x_n; t_1, \ldots, t_n) = f(x_1, \ldots, x_n; t_1 + c, \ldots, t_n + c)$$

Definition (Entropy rate): The entropy of a stochastic process $\{X_i\}$ is

$$H(\mathcal{X}) = \lim_{n \to \infty} \frac{1}{n} H(X_1, \ldots, X_n)$$

when the limit exists. Another definition for entropy rate is

$$H'(\mathcal{X}) = \lim_{n \to \infty} H(X_n | X_1, \ldots, X_{n-1})$$

when the limit exists. Note that the first denotes the per symbol entropy and the second one denotes the entropy of the last symbol given its past.

Theorem 3.4. *For discrete stationary stochastic processes both $H(\mathcal{X})$ and $H'(\mathcal{X})$ exist and they are equal.*

Proof. Since X_i are stationary then $b_n = H(X_n | X_1, \ldots, X_{n-1})$

$$H(X_{n+1} | X_1, \ldots, X_n) \leq H(X_{n+1} | X_2, \ldots, X_n) = H(X_n | X_1, \ldots, X_{n-1})$$

is non-increasing and bounded from below by 0 therefore, it converges. Moreover,

$$\begin{aligned} \frac{1}{n} H(X_1, \ldots, X_n) &= \frac{1}{n} [H(X_1) + H(X_2 | X_1) + \cdots + H(X_n | X_1, \ldots, X_{n-1})] \\ &= \frac{1}{n} (b_1 + \cdots + b_n) \to \inf b_n \\ \implies H(\mathcal{X}) &= H'(\mathcal{X}) \end{aligned}$$

Therefore, for stationary stochastic processes we can show the AEP

$$-\frac{1}{n} \lg f(X_1, \ldots, X_n) \xrightarrow{\mathbb{P}} H(\mathcal{X})$$

and from this we can define typical set for such processes, which has a size of $2^{nH(\mathcal{X})}$ and probability of close to 1. Then, we can show that there is code which one average uses $H(\mathcal{X})$ bits.