Notes for EECS 550: Information Theory

Yiwei Fu, Instructor: David Neuhoff

FA 2022

Contents

1 Source Codes **1**
1.1 Lossless Coding 1
1.1.1 Fixed-length to Fixed-length Block Codes (FFB) 2
1.2 Shannon-McMillian Theorem 7
1.3 Fixed Length to Variable Length (FVB) Lossless Source codes 9
1.4 Huffman's Code Design 12
1.5 Buffering 13
1.6 Universal Source Coding 14

2 Entropy **15**
2.1 Entropy 15
2.2 Basic Properties of Entropy 18
2.3 Conditional Entropy 20
2.4 Convexity 21

3 Information **22**
3.1 Information 22
3.2 Conditional Information 24
3.3 Cryptography From Information Perspective 24

4 Continuous Random Variables **25**
4.1 25

Office hours:

Chapter 1

Source Codes

1.1 Lossless Coding

Lossless coding is a type of data compression.

GOAL to encode data into bits so that

1. bits can be decoded perfectly or with very high accuracy back into original data;
2. we use as few bits as possible.

We need to model for data, a measure of decoding accuracy, a measure of compactness.

MODEL FOR DATA

Definition 1.1.1. A *source* is a sequence of i.i.d (discrete) random variables $U_1, U_2, \ldots$

We would like to assume a known alphabet $A = \{a_1, a_2, \ldots, a_Q\}$ and known probability distribution either through probability mass functions $p_U(u) = \Pr[U = u]$.

Definition 1.1.2. Source coding

PERFORMANCE MEASURES A measure of compactness (efficiency)

Definition 1.1.3. *Encoding rate*, also called *rate*, is the average number of encoded bits per data symbol.

There are two versions of average rate:

1. Emprical average rate

$$\langle r \rangle := \lim_{N \to \infty} \frac{1}{N} \sum_{k=1}^{N} L_k(U_1, \ldots, U_k),$$

2. Statistical average rate

$$\overline{r} := \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} \mathbb{E}[L_k(U_1, \ldots, U_k)]$$

where L_K is the number of bits out of the encoder after U_k and before U_{k+1}.

Definition 1.1.4. The per-letter frequency of error is defined as

$$\langle F_{LE} \rangle := \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} I(\hat{U}_k = U_k)$$

and per-letter error probability is defined as

$$p_{LE} := \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} \mathbb{E}[I(\hat{U}_k = U_k)] = \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} \Pr(\hat{U}_k = U_k)$$

1.1.1 Fixed-length to Fixed-length Block Codes (FFB)

characteristics

Definition 1.1.5. A code is *perfectly lossless* (PL) if the $\beta(\alpha(\underline{u})) = \underline{u}$ for all $\underline{u} \in A_U^k$ (the set of all sequences $u_1, \ldots, u_k$).

In order to be perfectly loss, α must be one-to-one. Encode must assign a distinct codeword (L bits) to each data sequences. rate = L/K. We seek $R^*_{PL}(k)$ the smallest rate of any PL code.

Number of sequences of size $k = Q^k$, and number binary sequence of size $L = 2^L$. We need $2^L \gg Q^K$.

$$\overline{r} = \frac{L}{k} \geq \frac{k \log_2 Q}{k} = \log_2 Q$$

Choose $\lceil k \log_2 Q \rceil$, then we have

$$R^*_{PL}(k) = \frac{\lceil k \log_2 Q \rceil}{k} \leq \frac{k \log_2 Q + 1}{k} = \log_2 Q + \frac{1}{k}.$$

$$\log_2 Q \leq R^*_{PL}(k) \leq \log_2 Q + \frac{1}{k}$$

Let R^*_{PL} be the least rate of any PL FFB code with any k. $R^*_{PL}(k) \to \log_2 Q$ as $k \to \infty$.

$R^*_{PL} = \inf_k R^*_{PL}(k)$

Now we want rate less and $\log_2 Q$ almost lossless codes.

$R^*_{AL} = \inf\{r,$ there is an FFB code with $\bar{r} \le r$ and arbitrarily small $P_{LE}\}$

$= \inf\{r,$ there is an FFB code with $\bar{r} \le r$ and $P_{LE} < \delta$ for all $\delta > 0\}$

Instead of per-letter probability P_{LE}, we focus on block error probability $P_{BE} = \Pr(\hat{\underline{U}} \neq \underline{U})$

Lemma 1.1.1. $P_{BE} \ge P_{LE} \ge \frac{P_{BE}}{k}$

Proof. See homework. ■

To analyze, we focus on the set of correctly encoded sequences. $G = \{\underline{u} : \beta(\alpha(\underline{u})) = \underline{u}\}$

Then we have

$$P_{BE} = 1 - \Pr[U \in G], |G| \le 2^k, L \ge \lceil \log_2 |G| \rceil .$$

<u>QUESTION</u> How large is the smallest set of sequences with length k form A_U with probability ≈ 1?

We need to use weak law of large numbers (WLLN).

Theorem 1.1.1. *Suppose $A_x = \{1, 2, \ldots, Q\}$ with probability $p_1, \ldots, p_Q$. Given $\underline{u} = (u_1, \ldots, u_k) \in A_U^k$.*

$$n_q(\underline{u}) := \#\textit{times } a_q \textit{ occurs in } \underline{u}, \quad f_q(\underline{u}) = \frac{n_q(\underline{u})}{k} = \textit{frequency}$$

Fix any $\varepsilon > 0$,

$$\Pr[f_q(\underline{u}) \doteq p_q \pm \varepsilon] \to 1 \textit{ as } k \to \infty.$$

Moreover,

$$\Pr[f_q(\underline{u}) \doteq p_q \pm \varepsilon, q = 1, \ldots, Q] \to q \textit{ as } k \to \infty.$$

<u>NOTATION</u> $a \doteq b \pm \varepsilon \iff |a - b| \le \varepsilon$

Consider subset of A_U^k that corresponds to this event x.

$T_k = \{\underline{u} : f_q(\underline{u}) \doteq p_q \pm \varepsilon, q = 1, \ldots, Q\}$.

$\Pr[\underline{U} = \underline{u}] = p(u_1)p(u_2)\ldots p(u_k)$.

By WLLN, $\Pr(T_k) \to 1$ as $k \to \infty$.

<u>KEY FACT</u> all sequences in T_k have approximately the same probability.

For $\underline{u} \in T_k$,

$$\begin{aligned} p(\underline{u}) &= p(u_1)p(u_2)\dots p(u_k) \\ &= p_1^{n_1(u)} p_2^{n_2(u)} \dots p_k^{n_k(u)} \\ &= p_1^{kf_1(u)} p_2^{kf_2(u)} \dots p_k^{kf_k(u)} \\ &\approx \tilde{p}^k \text{ where } \tilde{p} = p_1^{p_1} p_2^{p_2} \dots p_Q^{p_Q}. \end{aligned}$$

So we have $|T_k| \approx \frac{1}{\tilde{p}^k}$.

Then we have

$$\overline{r} = \frac{\log_2 |T_k|}{k} = -\frac{k \log_2 \tilde{p}}{k} = -\log_2 \tilde{p}.$$

Is that rate good? Can we do better? Can we have a set S with probability ≈ 1 and significantly smaller?

Since $\Pr(\underline{U} \in A_U^k \setminus T_k) \approx 0 \implies \Pr(\underline{U} \in S) \approx \Pr(\underline{U} \in S \cap T_k) \approx \frac{|S|}{|T_k|}$. So when k is large, T_k is the smallest set with large probability. And $R_{AL}^* \approx -\log \tilde{p}$.

How to express $\tilde{p}$.

$$\begin{aligned} -\log \tilde{p} &= -\log \prod_{i=1}^{Q} p_i^{p_i} \\ &= -\sum_{i=1}^{Q} p_i \log p_i =: \text{entropy} = H. \end{aligned}$$

Some properties of H:

1. its unit is bits
2. $H \geq 0$.
3. $H = 0 \iff p_q = 1$ for some q.
4. $H \leq \log_2 Q$.
5. $H = \log_2 Q \iff p_q = \frac{1}{Q}$ for all q.

Identify the set that WLLN says has probability $\to 1$. Suppose $X_1, X_2, \dots$ i.i.d. real-valued variables.

$$T = \{\underbrace{x_1 \dots x_n}_{\underline{x}} \in A_X^N : \frac{1}{N}\sum_{i=1}^{N} x_i \doteq \overline{x} \pm \varepsilon\}$$

is called a typical set. $\Pr(\underline{X} \in T) \approx 1$ when N is large.

Now suppose $X_1, X_2, \dots$ i.i.d. A_x-valued random variables, function $g : A_x \to \mathbb{R}$. Consider $Y_1, Y_2, \ldots$ with $Y_i = g(X_i)$. Y_i's are i.i.d. random variables.

If $\mathbb{E}[g(X)]$ is finite than we can apply WLLN that

$$\Pr\left(\frac{1}{N}\sum_{i=1}^{N} Y_i \doteq \mathbb{E}[Y] \pm \varepsilon\right) \to 1 \quad \text{as} \quad N \to \infty.$$

Typical sequences wrt g:

$$T^N_{x,p_\lambda,g,\varepsilon} = \left\{\underline{x} : \frac{1}{N}\sum_{i=1}^{N} g(x_i) \doteq \overline{g(X)} \pm \varepsilon\right\}.$$

If $\mathbb{E}[g(X)]$ is finite then by WLLN we have

$$\Pr(\underline{X} \in T_g) \to 1 \quad \text{as} \quad N \to \infty.$$

Example 1.1.1 (Indicator function)**.** Suppose $F \subset A_X$, and $g(x) = \begin{cases} 1 & x \in F \\ 0 & x \notin F \end{cases}$. Then $\frac{1}{N}\sum_{i=1}^{N} g(x_i) = f_F(x)$. Now

$$T_g = \{\underline{x} : f_F(x) \doteq \Pr(X \in F) \pm \varepsilon\}.$$

By WLLN,

$$\Pr(\underline{X} \in T_g) \to 1 \quad \text{as} \quad N \to \infty, \implies \Pr(n_F(\underline{X}) \doteq \mathbb{E}[x] \pm \varepsilon) \to 1.$$

Example 1.1.2. $A_x = \mathbb{R}, g(x) = x^2$. $T_g = \{\underline{x} :\}$

Theorem 1.1.2. *Now suppose M functions $g_1, g_2, \ldots, g_M$. Fix ε. Then*

$$T_{g_1,g_2,\ldots,g_M} = \bigcap_{i=1}^{M} T_{g_i}.$$

$$\Pr(\underline{X} \in T_{g_1,g_2,\ldots,g_M}) \to 1 \quad \textit{as} \quad N \to \infty.$$

Proof.

$$\begin{aligned}
\Pr(\underline{X} \notin T_{g_1,g_2,\dots,g_M}) &= \Pr\left(\underline{X} \in \left(\bigcap_{i=1}^{M} T_{g_i}\right)^c\right) \\
&= \Pr\left(\underline{X} \in \left(\bigcup_{i=1}^{M} T_{g_i}^c\right)\right) \\
&\le \sum_{i=1}^{M} \Pr(\underline{X} \in T_{g_i}^c) \to 0 \quad \text{as} \quad N \to \infty.
\end{aligned}$$

■

IMPORTANT APPLICATION

Suppose $A_x = \{a_1, \dots, a_Q\}$ a finite alphabet with probability $p_1, \dots, p_Q$. The $g_q(x)$ be the indicator of a_q. $T_q = \{\underline{x} : f_q(\underline{x}) \doteq p_q \pm \varepsilon\}$. And $\tilde{T} = \bigcap_{i=1}^{Q} T_i = \{\underline{x} : \forall q, f_q(x) \doteq p_q \pm \varepsilon\}$. $\tilde{T}^N_{X,p_X,\varepsilon}$ very typical sequence. We have

$$\Pr(\underline{X} \in \tilde{T}) \to 1 \quad \text{as} \quad N \to \infty.$$

If $\underline{x} \in \tilde{T}$, then $\underline{x} \in \tilde{T}_g$ for any other g. Consider any real-valued g. If $\underline{x} \in \tilde{T}_\varepsilon$ then $\underline{x} \in T_{g,\varepsilon c}$ for some c.

$$\frac{1}{N}\sum_{i=1}^{N} g(x_i) = \sum_{q=1}^{Q} \frac{n_q(x)}{N} g(Q_q) = \sum_{q=1}^{Q} (p_q \pm \varepsilon) q(a_q) = \mathbb{E}[g(X)] + \varepsilon \sum_{q=1}^{Q} g(Q_q)$$

$\Pr(\underline{X} \in \tilde{T}) \to 1$ as $N \to \infty$.

If $\underline{x} \in \tilde{T}$,

$$\begin{aligned}
p(\underline{x}) &= p(x_1)p(x_2)\dots p(x_N) \\
&= p_1^{n_1(\underline{x})} \dots \\
&= p_1^{f_1(\underline{x})N} \dots \\
&\doteq p_1^{(p_1 \pm \varepsilon)N} \dots \\
&\doteq 2^{N\left(\sum_{q=1}^{Q} p_q \log p_q \pm \varepsilon \sum_{q=1}^{Q} \log p_q\right)} \qquad \doteq 2^{-NH \pm N\varepsilon c}
\end{aligned}$$

Theorem 1.1.3 (Shannon-McMillian Theorem)**.** *Suppose $X_1, X_2, \dots$ i.i.d, $A_x = \{a_1, \dots, a_Q\}$ with probability $p_1, \dots, p_Q$. Then*

1.

$$\Pr(\tilde{X} \in \tilde{T}^N_\varepsilon) \to 1 \text{ as } N \to \infty.$$

2. *If* $\underline{x} \in \tilde{T}_\varepsilon^N$, $p(\underline{x}) \doteq 2^{-NH \pm N\varepsilon c}$.
3. $\left|\tilde{T}_\varepsilon^N\right| \doteq \Pr(\underline{X} \in \tilde{T}_\varepsilon^N) 2^{N(H \pm \varepsilon x)}$.

Proof. ■

1.2 Shannon-McMillian Theorem

Is $\tilde{T}$ essentially the smallest set with probability ≈ 1?

Yes. Let $S \in A_x^N$. We have

$\Pr(\underline{X} \in S = \Pr(X \in S \cap \tilde{T}) + \Pr(X \in S \cap \tilde{T}^c) \doteq |S \cap \tilde{T}| 2^{-NH \pm 2N\varepsilon c} + \Pr(\tilde{T}^c) \to 0$ as $N \to \infty$.

Theorem 1.2.1. *For every* $\varepsilon > 0$*, there is a sequence* $b_{\varepsilon,1}, b_{\varepsilon,2}, \ldots$ *s.t.* $b_{\varepsilon,N} \to 0$ *as* $N \to \infty$*,* $b_{\varepsilon,B} \geq 0$.

For any N *and any* $S \subset A_X^N$*,*

$$|S| \geq \left(\Pr(\underline{X} \in S) - b_{\varepsilon,N}\right) 2^{NH - N\varepsilon c}.$$

An in hindsight shortcut

Let us directly consider

$$\begin{aligned} T_{S,\varepsilon}^N &= \left\{\underline{x} : p(\underline{x}) \doteq 2^{-N(H \pm \varepsilon)}\right\} \\ &= \left\{\underline{x} : -\frac{1}{N} \log p(\underline{x}) \doteq H \pm \varepsilon\right\} \\ &= \left\{\underline{x} : -\frac{1}{N} \sum_{i=1}^{N} \log p(x_i) \doteq H \pm \varepsilon\right\} \end{aligned}$$

compare $\tilde{T}_\varepsilon^N$ and $T_{s,\varepsilon}^N$.

Claim: $\tilde{T}_\varepsilon^N \subset T_{s,\varepsilon}^N$ where $c = -; \sum_{q=1}^{Q} \log p_q$.

Suppose $\underline{x} \in \tilde{T}_\varepsilon^N$. Show if it is also in $T_{s,\varepsilon}^N$. Check the following $p(x) \doteq 2^{-NH \pm N\varepsilon c}$, $-\log p(x) \doteq NH \pm N\varepsilon c$.

$$\begin{aligned}
-\log p(\underline{x}) &= -\log \prod_{i=1}^{N} p(x_i) \\
&= -\log \prod_{q=1}^{Q} p_q^{n_q(x)} \\
&= -\log \prod_{q=1}^{Q} p_q^{N f_q(x)} \\
&\doteq -\log \prod_{q=1}^{Q} p_q^{N(p_q \pm \varepsilon)} \\
&\doteq -\sum_{q=1}^{Q} N(p_q \pm \varepsilon) \log p_a \\
&\doteq NH \pm N\varepsilon \sum_{q=1}^{C} \log p_k \\
&\doteq NH \pm N\varepsilon c.
\end{aligned}$$

Extreme example:

$A_x = \{0, 1\}, p_0 = p_1 = \frac{1}{2}$. $H = 1$.

$p(\underline{x}) = 2^{-N}$.

$T_{s,\varepsilon}^N = \left\{\underline{x} : p(\underline{x}) = 2^{-N(H\pm\varepsilon)} = 2^{-N}\right\} = A_X^N$.

$\tilde{T}_{\varepsilon}^N = \left\{\underline{x} : n_1(\underline{x} \doteq N\left(\frac{1}{2} + \varepsilon\right))\right\}$.

$|T_{s,\varepsilon}^N| \ddot{=} 2^{N(H\pm\varepsilon)}, |\tilde{T}_{\varepsilon}^N| \ddot{=} 2^{N(H\pm 2\varepsilon c)}$.

T_s is called probability typical. $\tilde{T}$ is called frequency typical.

Example $A_x = \{0, 1\}$, $p_1 = \frac{1}{4}, p_0 = \frac{3}{4}$. $\tilde{T}_{\varepsilon}^N = \left\{\underline{x} : f_1(\underline{x}) \doteq \frac{1}{4} + \varepsilon\right\}$.

$T_{s,\varepsilon}^T = \left\{\underline{x} : f_1(\underline{x}) = \frac{1}{4} \pm N\varepsilon \log \frac{1-p_1}{p_1}\right\}$

Typical sequences for an infinite alphabet

There are two cases: A_x is countably infinite / random variables are continuous

In the first case, frequency typical approach doesn't work. Probabilistic typical approach works just as is. $H = -\sum_{q=1}^{\infty} p_q \log p_q$ can be infinite.

Let $S_{\delta,N}$ = size of the smallest set of N sequences form A_x with probability at least $1-\delta$. Then for any $0 < \delta < 1$ and any h, $\frac{S_{\delta,N}}{2^N h} \to \infty$ as $N \to \infty$.

1.3 Fixed Length to Variable Length (FVB) Lossless Source codes

Recall that FFB perfectly lossless has $R^*_{PL} = \log_2 |A_x|$, and FFB almost lossless has $R^*_{AL} = H$.

FVB perfectly lossless $R^*_{VL} \leq \log_2 |A_x|$.

Suppose we have a source with $A_x = \{a, b, c, d\}$ with probability $\left\{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right\}$.

$p(u)$	u	code1	code2	code3	code4	code5	code6
$\frac{1}{2}$	a	00	0	0	0	0	0
$\frac{1}{4}$	b	01	10	10	10	1	01
$\frac{1}{8}$	c	10	110	10	11	01	011
$\frac{1}{8}$	d	11	111	11	111	10	0111
	Rate	2	1.75	1.5	1.625	1.25	1.875

We can see that code 3-5 are all bad.

Code 6 has an advantage that you know 0 represents the start of a codeword. We will see later why (Example 1.4.2).

FVB source code is characterized by

- source length k
- codebook of binary codewords $C = \left\{\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_{Q^K}\right\}$, $Q = |A_U|$.
- encoding rule $\alpha : A_U^K \to C$
- decoding rule $\beta : C \to A_U^K$.

The encoder operates in block fashion. The decoder does not.

Distinguish codes that look like code2 and codes that look like code6.

Definition 1.3.1. A codebook C is *prefix-free* if no codeword is the prefix of another.

A prefix-free code is called a prefix code. We will stick to prefix codes until states otherwise. (instantaneously decodable)

We like to draw binary tree diagrams of code.

Code 1:

233

A prefix is perfectly lossless if and only if α is 1-to-1. The rate: $\overline{r}(c) = \frac{\overline{L}}{K} = \frac{1}{K} \sum_{\underline{u}} p(\underline{u}) L(\underline{u})$

(length of codeword assigned to $\underline{u}$)

$$R_{VL}^*(k) = \min\left\{\overline{r}(c) : c \text{ is perfectly lossless FVB with source length } k\right\}.$$

$$R_{VL}^* = \inf\left\{\overline{r}(c) : c \text{ is PL FVB prefix code with any source length}\right\} = \inf_K R_{VL}^*(k).$$

How does one design a prefix code to have small or smallest rate?

Focus first $k = 1$. Shannon's idea: $L_q \approx -\log_2 p_q$.

$\sum_{q=1}^Q p_q L_q \approx -\sum_{q=1}^Q p_q \log p_q = H$.

Question. Is there a prefix code with $L_q \approx -\log p_q$ for $q = 1, 2, \ldots, Q$? Could there be prefix codes with even smaller rate?

Theorem 1.3.1 (Kraft inequality theorem)**.** *There is a binary prefix code with length $L_1, L_2, \ldots, L_Q$ iff the Kraft sum*

$$\sum_{q=1}^Q 2^{-L_q} \le 1.$$

Proof. Suppose $\underline{v}_1, \ldots, \underline{v}_Q$ is a prefix code with length $L_1, \ldots, L_Q$. Let $L_{\max} = \max_q L_q$.

From the tree, the number of sequences of length $L_{\max}$ prefixed by any codeword, is $\sum_{q=1}^Q 2^{L_{\max} - L_q} \le 2^{L_{\max}} \implies \sum_{q=1}^Q 2^{-L_q} \le 1$. So the Kraft inequality holds. ■

Now suppose

$$L_q = \left\lceil -\log_2 p_q \right\rceil, q = 1, \ldots, Q. \tag{1.3.1}$$

Is there a code with these lengths? Check Kraft.

$$\begin{aligned}
\sum_{q=1}^Q 2^{-L_q} &= \sum_{q=1}^Q 2^{-\lceil -\log p_q \rceil} \\
&\le \sum_{q=1}^Q 2^{-(-\log p_q)} \\
&\le \sum_{q=1}^Q 2^{\log p_q} \\
&\le \sum_{q=1}^Q p_q = 1.
\end{aligned}$$

So the Kraft inequality holds. $\exists$ a prefix code with length $L_1, \ldots, L_Q$ given by (1.3.1), called Shannon-Fano code.

Now the question is how good is this Shannon-Fano Code?

For the Shannon-Fano code, the rate (average length) is

$$\overline{L}_{SF} = \sum_{q=1}^{Q} p_q \left\lceil -\log p_q \right\rceil .$$

We have the following bounds:

$$H = \sum_{q=1}^{Q} p_q(-\log p_q) \leq \overline{L}_{SF} < \sum_{q=1}^{Q} p_q(-\log p_q + 1) = H + 1.$$

Question. Can we do better now?

We will show that $\overline{L} \geq H$ for any prefix code.

Let C be a prefix code with length $L_1, \ldots, L_Q$. Take the difference $\overline{L} - H = \sum_{q=1}^{Q} p_q L_q + \sum_{q=1}^{Q} p_q \log p_q$.

$$\begin{aligned}
\overline{L} - H &= \sum_{q=1}^{Q} p_q L_q + \sum_{q=1}^{Q} p_q \log p_q \\
&= -\sum_q p_q \log \frac{2^{-L_q}}{p_q} \\
&= -\sum_q p_q \ln \frac{2^{-L_q}}{p_q} \frac{1}{\ln(2)} \\
&\geq -\sum_q p_q \left(\frac{2^{-L_q}}{p_q} - 1 \right) \frac{1}{\ln(2)} \\
&\geq -\frac{1}{\ln(2)} \sum_q 2^{-L_q} + \sum_q p_q \frac{1}{\ln(2)} = \frac{1}{\ln 2}(1-1) = 0.
\end{aligned}$$

In homework we will that that $\overline{L}$ can get very close to $H + 1$.

Now allow $k \geq 1$. WE have a $C = \left\{ \underline{v}_1, \ldots, \underline{v}_{Q^k} \right\}$ of length $L_1, \ldots, L_{Q^k}$. We want small

$$\overline{r}(c) = \frac{\overline{L}}{K} = \frac{\sum_u p(\underline{u}) L(\underline{u})}{k}.$$

Shannon-Fano code achieve that

$$H^k \leq \overline{L}_{SF} < H^k + 1 \implies \frac{H^k}{k} \leq \overline{r}_{SF} = \frac{\overline{L}_{SF}}{k} < \frac{H^k}{k} + \frac{1}{k}.$$

Since $H^k = kH$ we have

$$H \leq \overline{r}_{SF} < H^k + \frac{1}{k}.$$

Similarly we have for any prefix code, we have

$$\overline{r} = \frac{\overline{L}}{K} \geq \frac{H^k}{k} = H.$$

This leads to a new coding theorem.

Theorem 1.3.2. *Given i.i.d. source U with alpha A_U and entropy H. We have*

1. *For every k, $R^*_{VL} \leq R^*_{VL}(k) < H + \frac{1}{k}$.*
2. *For every k, $R^*_{VL}(k) \geq R^*_{VL} \geq H$.*

Combined we have

$$\forall k \in \mathbb{Z}_{>0}, H \leq R^*_{VL}(k) < H + \frac{1}{k}$$

and

$$R^*_{VL} = H.$$

1.4 Huffman's Code Design

Given $p_1, \ldots, p_Q$, it finds a prefix code with smallest $\overline{L}$.

Algorithm 1.4.1 Huffman Code

Input: Alphabet probability $\{p_i | i = 1, \ldots, Q\}$, WLOG assume $p_1 \geq p_2 \geq \ldots \geq p_Q$.
Output: FVB Codebook for alphabet $\{a_i | i = 1, \ldots, Q\}$.

function HUFFMAN($P_Q = \{p_i | i = 1, \ldots, Q\}$)
    if $Q = 2$ **then return** $\{0, 1\}$
    end if
    $p'_{Q-1} \leftarrow p_{Q-1} + p_Q$
    $P_{Q-1} \leftarrow (P_Q \setminus \{p_{Q-1} + p_Q\}) \cup \left\{p'_{Q-1}\right\}$
    $c_{Q-1} \leftarrow$ HUFFMAN$(P_{Q-1}) =: \left\{\underline{v}_1, \ldots, \underline{v}_{Q-1}\right\}$
    $c_Q \leftarrow \left\{\underline{v}_1, \ldots, \underline{v}_{Q-2}, \underline{v}_{Q-1}0, \underline{v}_{Q-1}1\right\}$
end function

Proposition 1.4.1. *If c_{Q-1} is optimal for P_{Q-1} then c_Q is optimal for P_Q.*

Example 1.4.1.

We found that

But there is a tighter upper bound

$$\overline{L}^* \leq \begin{cases} H + p_{\max} & p_{\max} < \frac{1}{2} \\ H + p_{\max} + 0.086 & p_{\max} \geq \frac{1}{2}. \end{cases}$$

Hence

$$\mathscr{R}^*_{VL}(k) \leq \begin{cases} H + \frac{p^k_{\max}}{k} & (p_{\max})^k < \frac{1}{2} \\ H + \frac{p^k_{\max}}{k} + \frac{0.086}{k} & (p_{\max})^k \geq \frac{1}{2}. \end{cases}$$

Up till now we've only focused on i.i.d RV's. Now suppose RV's are dependent, then

$$\frac{H^k}{k} < \frac{kH}{k}.$$

For a stationary random process,

$$\frac{H^k}{k} \searrow H_\infty.$$

For example, English has

$$H^1 \approx 4.08, H_\infty \approx 1.$$

The bits produced by a good lossless source code ($\overline{r} \approx H$) are approximately i.i.d. equiprobable.

Synchronization and transmission entropy

Example 1.4.2. Suppose $\{01, 001, 101, 110\}$ for $\{a, b, c, d\}$.

$$\underline{u} = ddddddddd\ldots \implies \underline{z} = 110110110110110110\ldots$$
$$\text{(if one leading 1 is missing) } \underline{z}' = 101101101101\ldots \implies \hat{u} = ccccccc\ldots$$

Now if $\{1, 01, 001, 000\}$ for $\{a, b, c, d\}$. Then the same problem will not happen.

1.5 Buffering

Suppose the source is outputting at R symbols per second. The encoder would have $R\overline{r}$ bits per second.

Buffer overflow happens when a long sequence of low probability symbols are encoded.

Buffer underflow happens when a long sequence of high probability symbols are encoded. Buffer will be empty. Include an additional codeword in codebook called a "flag". Insert this codeword when buffer becomes empty.

We focused prefix codes. There are some non prefix codes that can be decoded losslessly.

Definition 1.5.1. A code is *separable* if any finite sequence of codewords is different from any other finite sequence of codewords.

Remark. Prefix codes are separable. And determining if a non prefix code is separable is not easy.

Could prefix codes have smaller I?

McMillian's theorem says that Kraft inequality holds for separable codes. If you have a separable code whose codeword satisfy Kraft, then there is a prefix code with same lengths.

Lossless coding for source with infinite alphabet.

Suppose $A_n = \{1, 2, 3, \ldots\}$.

1. FFB codes can't have finite rate if perfectly lossless
2. AL FFB then SM theorem
3. FVB. Current approach is based on Kraft inequality. It still holds for infinite case. (See Apppendix). But Huffman's optimal design does not apply.

Other forms of variable length lossless source codes.

1. Run-length coding
2. Dictionary coding

1.6 Universal Source Coding

Suppose you are to encode 10^6 symbols from alphabet $A = \{a, b, c, d\}$. We can calculate $n_a(\underline{u}), n_b(\underline{u}), n_c(\underline{u}), n_d(\underline{u})$ and similarly, frequencies. Then we can apply Huffman or Shannon-Fano code.

Chapter 2

Entropy

The star of this chapter is

$$\ln x \leq x - 1$$

2.1 Entropy

Entropy

$$H := -\sum_x p(x) \log p(x)$$

is a measure of randomness or uncertainty.

$$H_q := -\sum_x p(x) \log_q p(x), \quad H_q = H_r \frac{1}{\log_r q}.$$

$$H(X) = \mathbb{E}[\log p_X(X)]$$

Remark. 1. $H(X) \geq 0$ and $H(X) = 0 \iff p(x) = 1$ for some x.

2. $H(X) = \infty$ if X is continuous or has continuous component.
3. $H(X, Y) \geq H(X)$
4. $H(X, Y) = H(X) + H(Y)$ if X and Y are independent.
5. $H(X_1, X_2, \ldots, X_n) = H(X_{\sigma(1)}, \ldots, X_{\sigma(n)})$ for any $\sigma \in S_n$.

Divergence is a measure of dissimilarity of two probability distribution.

Definition 2.1.1. Suppose p and q are probability mass functions. The *divergence* from p

to q is

$$D(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

Remark. 1. $p = q \implies D(p\|q) = 0$

2. It is *not symmetric*. $D(p\|q) \neq D(q\|p)$. You can make symmetric by taking the sum, but then it is not nicely related to information theory.

What if $p(x) = 0$ for some x? We take $0 \log \frac{0}{q(x)} = 0$. So if $\exists x \; s.t. \; q(x) = 0$ and $p(x) \neq 0$. Then $D(p\|q) = \infty$.

When alphabet A_x is infinite, $D(p\|q)$ can be ∞ even when $p(x) > 0$ and $q(x) > 0$ for all $x \in A_x$.

Is $\sum_x p(x) \log \frac{p(x)}{q(x)}$ is always defined? Write

$$\sum_x p(x) \log \frac{p(x)}{q(x)} = \sum_{x, p(x) > q(x)} \log \frac{p(x)}{q(x)} + \sum_{x, p(x) < q(x)} \log \frac{p(x)}{q(x)}$$

We will show later that the second term is never $-\infty$, so it is always well-defined.

Proposition 2.1.1 (Divergence inequality)**.** *For any p, q,*

$$D(p\|q) \geq 0, D(p\|q) = 0 \iff p = q.$$

Proof.

$$\begin{aligned} D(p\|q) &= \sum_x p(x) \log \frac{p(x)}{q(x)} \\ &= \sum_x p(x) \ln \frac{p(x)}{q(x)} \frac{1}{\ln 2} \\ &= -\sum_x p(x) \ln \frac{q(x)}{p(x)} \frac{1}{\ln 2} \\ &\leq -\sum_x p(x) \frac{q(x) - p(x)}{p(x)} \frac{1}{\ln 2} \\ &\leq -\left(\sum_x p(x) - q(x) \right) \frac{1}{\ln 2} = 0 \end{aligned}$$

For first equality, $\impliedby$ is clear. Now suppose $D(p\|q) = 0$. Then

$$\ln \frac{q(x)}{p(x)} = \frac{q(x)}{p(x)} - 1 \implies p(x) = q(x) \text{for all } x \text{ with } p(x) > 0.$$

■

Let's rewrite the divergence inequality a little bit.

$$0 \le D(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = -H(x) - \sum_x p(x) \log(q_x)$$

$$H(x) \le -\sum_x p(x) \log(q_x)$$

with $= \iff p = q$. $-\sum_x p(x) \log(q_x)$ is called cross entropy.

Definition 2.1.2. The *cross entropy* of p with respect to q is

$$H_c(p, q) := -\sum_x p(x) \log q(x).$$

Cross-entropy inequality: for any p, q

$$H_p(X) \le H_c(p, q)$$

with $= \iff p = q$.

Remark.

$$D(p\|q) = H_c(p, q) - H_p(X) \iff H_c(p, q) = H_p(X) + D(p\|q).$$

Definition 2.1.3. Variation distance

$$V(p, q) = \sum_x |p(x) - q(x)|$$

How does $D(p\|q)$ compare to $V(p, q)$?

Proposition 2.1.2 (Pinskev's inequality)**.**

$$V(p, q) \le \sqrt{(2 \ln 2) D(p\|q)}.$$

So small $D(p\|q) \implies$ small $V(p, q)$ even when $|A_x| = \infty$. On the other hand the converse is not true.

Lemma 2.1.1. *If there exists $0 < \delta < 1$ such that $\frac{|p(x)-q(x)|}{p(x)} \le \delta$ for all x s.t. $p(x) > 0$ then*

$$D(p\|q) \le \frac{\delta}{1-\delta} \frac{1}{\ln 2}$$

If $D(p\|q) \approx 0$ then $p \approx q$, meaning $V(p, q) \approx 0$.

If p, q are percentage wise close, then $D(p\|q) \approx 0$.

Log-sum inequality

Suppose $u_1, u_2, \ldots, u_n$, $v_1, \ldots, v_n$ nonnegative. Then

$$\sum_i u_i \log \frac{u_i}{v_i} \geq \left(\sum_i u_i\right) \log \frac{\sum_i u_i}{\sum_i v_i}.$$

This is a generalization of divergence inequality.

2.2 Basic Properties of Entropy

Proposition 2.2.1.

$$H(X^N) \leq \sum_{i=1}^{N} H(X_i)$$

with $= \iff X_1, \ldots, X_N$ *are independent.*

Proof.

$$H(X^N) \leq H_L(p, q)$$

where p is probability mass function of $X_1, \ldots, X_N$. Choose

$$q(x_1, \ldots, x_n) = p(x_1)p(x_2)\ldots p(x_n).$$

■

Since we are dealing with discrete random variables, it is useful to think about probability mass function as a set of probabilities $\{p_1, p_2, \ldots\}$. Write

$$H(p_1, p_2, \ldots) = -\sum_i p_i \log p_i$$

Let $p_i' = p_i + p_j$, replace p_i, p_j with p_i' and leave all others the same. We have

$$-p_i \log p - i - p_j \log p_j \geq -(p_i + p_j)\log(p_i + p_j)$$

i.e. entropy decreases when two probabilities are merged.

Proposition 2.2.2. *If X is Q-ary with $Q < \infty$ then*

$$H(X) \leq \log_2 Q$$

Proof. Let $q(x) = \frac{1}{Q}$.

$$\begin{aligned} H(X) \leq H_c(p, q) &= -\sum_x p(x) \log p(x) \\ &= \sum_x p(x) \log_2 Q = \log_2 Q. \end{aligned}$$

■

Proposition 2.2.3. *Suppose $Y = g(X)$. Then*

$$H(Y) = H(g(X)) \leq H(X)$$

and $= \iff g$ is one-to-one (probabilistically).

Proof.

$$H(Y) = \sum_y p(y) \log p(y)$$

where $p(y) = \sum_{x, g(x) = y} p(x)$. ■

$p_i = q^i(1-q), i = 0, 1, 2, \ldots$, then

$$
\begin{aligned}
H(X) &= -\sum_{i=0}^{\infty} p_i \log p_i \\
&= -\sum_{i=0}^{\infty} q^i(1-q) \log q^i(1-q) \\
&= -\sum_{i=0}^{\infty} q^i(1-q) \left(i \log q + \log(1-q)\right) \\
&= -\log q \sum_{i=0}^{\infty} q^i(1-q)i - \sum_{i=0}^{\infty} q^i(1-q) \log(1-q) \\
&= -\log q \cdot q(1-q) \frac{\mathrm{d}}{\mathrm{d}q} \sum_{i=0}^{\infty} q^i - \log(1-q) \\
&= -\log q \cdot q(1-q) \frac{1}{(1-q)^2} - \log(1-q) \\
&= -\log q \cdot \frac{q}{1-q} - \log(1-q) = \frac{-q\log(q) - (1-q)\log(1-q)}{1-q} \\
&= H(Q) \frac{1}{1-q} < \infty.
\end{aligned}
$$

$p_i = \frac{\alpha}{i(\ln i)^2}, i = 2, 3, \ldots$, then

$$
\begin{aligned}
H(X) &= -\sum_{i=2}^{\infty} \frac{\alpha}{i(\ln i)^2} \log\left(\frac{\alpha}{i(\ln i)^2}\right) \\
&=
\end{aligned}
$$

2.3 Conditional Entropy

$$
H(X \mid Y) = \sum_{x,y} p(x,y) \log p(x|y) \geq 0
$$

with equality iff X is a function of Y.

$$
H(X \mid Y) \leq H(X)
$$

with equality iff X, Y are independent.

chain rule:

$$
H(X,Y) = H(X) + H(Y \mid X) \implies H(X,Y) \geq H(X)
$$

Conditional lossless source coding

2.4 Convexity

Goal: entropy is a concave (convex $\cap$).

Extended definition of entropy

Definition 2.4.1.

$$\overline{H}(x) = \sup_{\text{finite quantizers } Q} H(Q(x))$$

where finite quantizer is a function $Q : A \to B, |\{Q(x) : x \in A\}| < \infty$.

This gives normal definition for discrete random variables and ∞ for continuous and mixed random variables

Chapter 3

Information

3.1 Information

Not a good question: How much information is there in X? Better questions: The information in X about random variable Y?

Definition 3.1.1. The (mutual) information given by X about Y is defined as

$$I(X;Y) := H(Y) - H(Y \mid X)$$

Definition 3.1.2. Y = outcome of a fair 6-sided die. X = oddity of the outcome. We have

$$H(Y) = \log_2 6, H(Y \mid X) = \log_2 3 \implies I(X;Y) = \log_2 6 - \log_2 3 = \log_2 2 = 1.$$

Lemma 3.1.1. *Suppose X, Y are discrete random variables. Then*

1. $I(X;Y) \geq 0$, $= 0$ *iff* X, Y *independent.*

2. *We have alternate formulas*

$$\begin{aligned}
I(X;Y) &= -\sum_y p(y)\log p(y) + \sum_{x,y} p(x,y)\log p(y|x) \\
&= -\sum_{x,y} p(x,y)\log p(y) + \sum_{x,y} p(x,y)\log p(y|x) \\
&= \sum_{x,y} p(x,y)\log \frac{p(y|x)}{p(y)} \\
&= \sum_{x,y} p(x,y)\log \frac{p(x,y)}{p(x)p(y)}
\end{aligned}$$

This shows that I is symmetric: $I(X;Y) = I(Y;X)$.

3. *From above,*

$$\begin{aligned}
I(X;Y) = &-\sum p(x,y)\log p(x) - \sum p(x,y)\log p(y) + \sum p(x,y)\log p(x,y) \\
&= H(X) + H(Y) - H(X,Y).
\end{aligned}$$

4. *We can view information as an expectation:*

$$I(X;Y) = \mathbb{E}\left[\log \frac{p(Y \mid X)}{p(Y)}\right] = \mathbb{E}\left[\log \frac{p(X \mid Y)}{p(X)}\right] = \mathbb{E}\left[\log \frac{p(X,Y)}{p(X)p(Y)}\right].$$

5. *We can view information with respect to divergence:*

$$I(X,Y) = D(p_{XY} \| p_X p_Y).$$

Remark. Alternate notation $I_\phi(X,Y), I_{X;Y}(p), I(p)$.

What happens if $p(x,y) = 0$, or $p(x) = 0$, or $p(y) = 0$? $p(x) = 0$ or $p(y) = 0 \implies p(x,y) = 0$.

Remark. $I(X;Y)$ is possible. Suppose $H(Y) = \infty$ and Y is a function of X. Then $I(X;Y) = H(Y) - H(Y \mid X) = \infty - 0 = \infty$.

Information for more variables

$$I(X,Y;V,W,Z) = H(X,Y) - H(X,Y \mid V,W,Z)$$

Relations between information entropy

1. $I(X;Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$.

2. $I(X;Y) \leq H(X)$, $=$ iff X is a function of Y.
3. $I(X;X) = H(X)$.
4. $I(X;g(X)) = H(g(X))$.

3.2 Conditional Information

Recall that we have two concepts of conditional entropy.

$$H(X \mid Y = y), H(X \mid Y) = \sum_y p(y) H(X \mid Y = y)$$

We are going to use the same approach for conditional information.

Definition 3.2.1. Suppose X, Y, Z are two discrete random variables,

$$I(X;Y|Z = z) = \sum_{x,y} p(x,y) \log \frac{p(xy \mid Z = z)}{p(x|Z = z)p(y|Z = z)}.$$

$$I(X;Y \mid Z) = \sum_z p(z) I(X;Y \mid Z = z).$$

Lemma 3.2.1. *1.* $I(X;Y \mid Z = z) \geq 0$

2. $I(X;Y \mid Z) \geq 0$, $= 0$ *iff* X, Y *are conditionally independent given* Z.

3. $I(X;Y \mid Z) = H(X \mid Z) - H(X \mid Y, Z)$.

4. $I(X;Y,Z) = I(X;Z) + I(X;Y \mid Z)$

chain rule of conditional information

$$I(X;YZ \mid U) = I(X;Z \mid U) + I(X;Y \mid Z, U).$$

3.3 Cryptography From Information Perspective

$A_X = A_K$ and $|A_X| = |A_K| = 2^N$. $p_K(k) = 2^{-N}, k \in A_K$.

If $|K| < |A_X|$ then the crypto system is not perfect.

Fix $x \in |A_X|$. For each y we have $P(Y = y|X = x) = P(Y = y) > 0$. Therefore for each y, there must be some key $k \in K$ such that $y = e_K(x)$. It follows that $|K| \geq |Y|$. The encryption is injective giving $|Y| \geq |A_X|$.

Chapter 4

Continuous Random Variables

4.1

Definition 4.1.1.

We assume alphabet is $\mathbb{R}$, X is absolutely continuous

Example 4.1.1. 1. Uniform

$$p(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

2. Gaussian

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

3. Laplacian

$$p(x) = \frac{1}{\sqrt{2}\sigma} e^{-\frac{\sqrt{2}}{\sigma}|x|}$$

4. Exponential

$$p(x) = \begin{cases} \frac{\sqrt{2}}{\sigma} e^{-\frac{\sqrt{2}}{\sigma}|x|} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Definition 4.1.2. The *support* of a random variable X or of its probablity distribution is defined by

$$S := \{x : \Pr(X \doteq x \pm \varepsilon) > 0, \forall \varepsilon > 0\}.$$

Definition 4.1.3. Conditional probability

$$\Pr(F \mid X = x) = \frac{\Pr(F, X = x)}{\Pr(X = x)}.$$

When X is continuous,

$$\Pr(F \mid X = x) = \lim_{\delta \to 0} \Pr(F \mid X \doteq x \pm \delta)$$

Suppose $Y = 3X$

$$1 = \Pr(Y = 3 \mid X = 1) = \lim_{\delta = 0} \frac{\Pr(Y = 3 \mid X \doteq 1 \pm \delta)}{\Pr(X \doteq 1 \pm \delta)} = 0, \text{ contradition}$$

So we use

$$\Pr(Y \in F \mid X = x) = \lim_{\varepsilon \to 0} \lim_{\delta \to 0} \Pr(Y \in F_\delta \mid X \doteq x \pm \varepsilon)$$

where

$$F_\delta = \{y \mid \|y - y'\| \text{ for some } y' \in F\}.$$

Generalized sum