Notes for EECS 550: Information Theory

Yiwei Fu, Instructor: David Neuhoff

FA 2022

Contents

Office hours:

Chapter 1

Introduction

1.1 What is Information Theory

1.2 Lossless Coding

It 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 1.2.1. 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 1.2.2. Source coding

PERFORMANCE MEASURES A measure of compactness (efficiency)

Definition 1.2.3. rate = encoding rate = average number of encoded bits per data symbol

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

Statistical avg rate:

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

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

Accuracy per-letter frequency of error

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

per-letter error probability

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

Fixed-length to fixed-length block codes (FFB)

characteristics

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 $u_1, \dots, u_k$).

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.

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

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

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

Lemma 1.2.1. $P_{BE} \geq P_{LE} \geq \frac{P_{BE}}{k}$

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

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

<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 1.2.1. *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> $a \doteq b \pm \varepsilon \iff |a - b| \leq \varepsilon$

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. $H = 0 \implies \iff p_q = 1$ 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 1.2.1 (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 1.2.2. $A_x = \mathbb{R}, g(x) = x^2$. $T_g = \{\underline{x} :\}$

Theorem 1.2.2. *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.

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

■

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 1.2.3 (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. ■

1.3

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

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

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

Theorem 1.3.1. *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* $S \subset A_X^N$*,*

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

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:

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

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

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

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

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

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 $N \to \infty$.

1.4 Perfectly Loss fixed length to variable length (FVB) lossless source codes

RECALL: FFB perfectly lossless $R^*_{PL} = \log_2 |A_x|$

FFB almost lossless $R^*_{AL} = H$

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.

Suppose 101110101110

Let say if one bit is changed to zero, 101010101110

Code 6 has an advantage that you know 0 represents the start of a codeword.

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 1.4.1. 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 $\underline{u}$)

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

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

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

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

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

Q: 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?

Kraft inequality theorem

Theorem 1.4.1. *There is a binary prefix code with length $L_1, L_2, \ldots, L_Q$ iff the Kraft sum*

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

$$L_q = \lceil -\log_2 p_q \rceil, q = 1, \ldots, Q. \tag{1.4.1}$$

Is there a code with these length? Check Kraft.

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

So the Kraft inequality holds. $\exists$ a prefix code with length $L_1, \ldots, L_Q$ given by (1.4.1), 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:

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

Can we do better now?

Will show $\overline{L} \geq H$ for any prefix code.

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

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