Shannon's random coding bound in terms of Information Density

Amizhthni PRK

April 18, 2024

Outline

- Why Information Density
- What is Information Density
- Encoder and Decoder Design
- Properties of Information Density
- Shannon's random coding bound using Information Density
- Channels and Channel Capacity
- Shannon's noisy channel coding theorem

We are tasked with the design of the encoder $f : [\mathrm{M}] \to \mathcal{X}$ or the Codebook $c_i \triangleq f(i), i \in [M]$

M-ary Hypothesis Testing

- We apply standard statistical tests. We consider M different Hypotheses i.e we test with respect all input words $\in [M]$
- $P_{X|c_1}, ..., P_{X|c_M}$
- Maximum testability, Optimal

Hard to optimise the Encoder Function & Less efficient. Is there a better way to choose a good f?

Suboptimal way

- Approximate the problem to M instances of a Binary Hypothesis Testing
- We test $P_{Y|X=c_m}$ against $P_Y \ \forall m \in [M]$
- P_Y is the average of all the probabilities conditioned over different inputs that result in the particular output Y.

This leads to the Log Likelihood test from Neymann Pearson theorem as shown below

$$log \frac{P_{Y|X=x}}{P_Y}$$

The above intuition leads us to define a quantity termed as **Information Density**

Definition

Given joint distribution $P_{X,Y}$ we define information density $i_{P_{XY}}(x;y)$ as :

$$i_{P_{XY}}(x;y) = \log \frac{P_{Y|X}(y|x)}{P_Y(y)}$$

(explains the symmetric nature of Information Density)

$$i_{P_{XY}}(x;y) = \log \frac{P_{Y,X}(x,y)}{P_Y(y)P_X(x)}$$

Intuition:
For a given measurement of Y, what is the likelihood that the input was X.

We begin with the definition of a code.

Definition

An M-code for $P_{Y|X}$ is an encoder/decoder pair (f, g) of functions.
- encoder $f : [\mathrm{M}] \to \mathcal{X}$
- decoder $g : \mathcal{Y} \to [M] \cup e$

- $\forall i \in [M] : c_i \triangleq f(i)$ are *codewords*.
- The collection $\mathcal{C} = c_1.....c_m$ is called a *codebook*.
- $\forall i \in [M] : D_i \triangleq g^{-1}(i)$ is the *decoding region* for i.

We chain three objects the message W, the encoder and the decoder together into the following markov chain,

$$W \xrightarrow{f} X \xrightarrow{P_{Y|X}} Y \xrightarrow{g} \hat{W}$$

Decoder independent of P_X.

Maximum likelihood decoder

$$g^*(y) = \underset{m\in[M]}{\text{argmax}}\, i(c_m; y)$$
$$= \underset{m\in[M]}{\text{argmax}}\, P_{Y|X}(y|c_m)$$
$$= \underset{m\in[M]}{\text{argmax}}\, P_{X|Y}(c_m|y)$$

$P_{Y|X}$ is fixed, becomes a function of only P_X. We make a justified choice of P_X and run binary tests by thresholding each information density $i(c_i; y)$.

Threshold Definition of Information Density

$$g(y) = \text{any } m \text{ s.t } i(c_m; y) > \gamma$$

Where γ is a threshold and P_X is chosen carefully.

Properties of Information Density

All standard properties of log-likelihood.

1. Expectation

The expectation $\mathbb{E}[i(X;Y)]$ is well defined and positive. We have,

$$I(X;Y) = \mathbb{E}[i(X;Y)]$$

Let $\bar{X} \perp\!\!\!\perp (X,Y)$ be a copy of X. The following properties hold.

1. Conditioning and unconditioning trick

- If $f(y) = 0$ and $g(x) = 0$ when $i(x;y) = -\infty$,

$$\mathbb{E}[f(Y)] = \mathbb{E}[\exp\{-i(x;Y)\}f(Y)|X = x] \ \forall x$$

$$\mathbb{E}[g(X)] = \mathbb{E}[\exp\{-i(X;y)\}g(X)|Y = y] \ \forall y$$

- If $f(x,y) = 0$ i(x;y)=-∞ ,

$$\mathbb{E}[f(\bar{X},Y)] = \mathbb{E}[\exp\{-i(X;Y)\}f(X,Y)]$$

$$\mathbb{E}[f(X,\bar{Y})] = \mathbb{E}[\exp\{-i(X;Y)\}f(X,Y)]$$

- For any function $f : \mathcal{X}x\mathcal{Y} \to \mathbb{R}$

$$\mathbb{E}[f(\bar{X}, Y)1\{i(\bar{X}; Y) > -\infty\}] = \mathbb{E}[f(X, Y)\exp\{-i(X; Y)\}]$$

- Let f_+ be a non-negative function. Then for P_X-almost every x we have

$$\mathbb{E}[f_+(\bar{X}, Y)1\{i(\bar{X}; Y) > -\infty\}|\bar{X} = x] = \mathbb{E}[f_+(X, Y)\exp\{-i(X; Y)\}|\bar{X} = x]$$

Corollary

For P_X-almost every x we have,

$$\mathbb{P}[i(x; Y) > t] \leq \exp(-t)$$

$$\mathbb{P}[i(\bar{X}; Y) > t] \leq \exp(-t)$$

Picking $f_+(x, y) = 1\{i(x; y) > t\}$ in the second result proves this.

Shannon's Achievability Bound

The main goal of error correcting code is to push different codewords as far as possible to minimize effects of channel noise.

Definition

Fix a channel $P_{Y|X}$ and an arbitrary input distribution P_X . Then for every $\tau > 0$ there exists an (M, *epsilon*)-code with

$$\epsilon \leq \mathbb{P}[i(X;Y) \leq \log M + \tau] + \exp(-\tau)$$

The proof is as follows,
In this bound we consider a threshold-based suboptimal decoder $g(y)$ as follows:

$$g(y) = \begin{cases} m, & \exists! c_m \text{ s.t. } i(c_m, y) \geq \log M + \tau \\ e, & \text{otherwise} \end{cases}$$

This decoder decodes g only when c_m is unique and has information density above the threshold $\log M + \tau$.
The error probability for this decoder is :

$$P_e(c_1, \ldots, c_M) = \mathbb{P}\left[\{i(c_W; Y) \leq \log M + \tau\} \cup \{\exists \bar{m} \neq W, i(c_{\bar{m}}; Y) > \log M + \tau\}\right]$$

The second step was the forgo the discrete optimization and instead generate the codebook $(c_1, ...c_m)$ randomly. We can then try reasoning about the mean. Which is calculated as :

$$
\begin{aligned}
&\mathbb{E}\left[P_e\left(c_1, \ldots, c_M\right)\right] \\
=&\mathbb{E}\left[P_e\left(c_1, \ldots, c_M\right) \mid W=1\right] \\
=&\mathbb{P}\left[\left\{i\left(c_1 ; Y\right) \leq \log M+\tau\right\} \cup\left\{\exists \bar{m} \neq 1, i\left(c_{\bar{m}}, Y\right)>\log M+\tau\right\} \mid W=1\right] \\
&\leq \mathbb{P}\left[i\left(c_1 ; Y\right) \leq \log M+\tau \mid W=1\right]+\sum_{\bar{m}=2}^{M} \mathbb{P}\left[i\left(c_{\bar{m}} ; Y\right)>\log M+\tau \mid W=1\right] \\
&\stackrel{(a)}{=} \mathbb{P}[i(X ; Y) \leq \log M+\tau]+(M-1) \mathbb{P}[i(\bar{X} ; Y)>\log M+\tau] \\
&\leq \mathbb{P}[i(X ; Y) \leq \log M+\tau]+(M-1) \exp(-(\log M+\tau)) \\
&\leq \mathbb{P}[i(X ; Y) \leq \log M+\tau]+\exp(-\tau)
\end{aligned}
$$

A crucial step (a) follows from the fact that given $W = 1$ and $\bar{m} \neq 1$,

$$(c_1, c_{\bar{m},Y}) \stackrel{(d)}{=} (X, \bar{X}, Y)$$

Since the average satisfies the given bound, there is atleast one $(c_1 ... c_m)$ satisfying the same bound.

- Channel : Fix an input alphabet $\mathcal{A}$ and an output alphabet $\mathcal{B}$. A sequence of Markov kernels $P_{Y^n|X^n} : \mathcal{A}^n \to \mathcal{B}^n$ indexed by the integer n = 1, 2 . . . is called a channel.
- A channel is memoryless if $P_{Y^n|X^n}$ factorizes into a product distribution. Namely,

$$P_{Y^n|X^n} = \prod_{k=1}^{n} P_{Y_k|X_k}.$$

  where each $P_{Y_k|X_k} : \mathcal{A} \to \mathcal{B}$; in particular, $P_{Y^n|X^n}$ are compatible at different blocklengths n.
- A channel is stationary memoryless if it is memoryless with $P_{Y_k|X_k}$ not depending on k, denoted commonly by $P_{Y|X}$. In other words,

$$P_{Y^n|X^n} = \left(P_{Y|X}\right)^{\otimes n}.$$

  Thus, in discrete cases, we have

$$P_{Y^n|X^n}\left(y^n \mid x^n\right) = \prod_{k=1}^{n} P_{Y|X}\left(y_i \mid x_i\right).$$

Notation

- An (n, M, ϵ)-code is an (M, ϵ)-code for $P_{Y^n|X^n}$, consisting of an encoder $f : [M] \to \mathcal{A}^n$ and a decoder $g : \mathcal{B} \to [M] \cup \{e\}$.
- An $(n, M, \epsilon)_{max}$-code is analogously defined for maximum probability of error.

The limits are naturally,

$$M^*(n, \epsilon) = \max\{M : \exists(n, M, \epsilon) - code\}$$

$$M^*(n, \epsilon)_{max} = \max\{M : \exists(n, M, \epsilon)_{max} - code\}$$

The *Transmission Rate* defined as $R = \frac{\log_2 M}{n}$ is the number of bits transmitted per channel use.
In this line it we study $\frac{1}{n} \log M^*(n, \epsilon)$.

Definition

The ϵ-capacity C_ϵ and Shannon capacity C are defined as follows

$$C_\epsilon \triangleq \liminf_{n\to\infty} \frac{1}{n} \log M^*(n,\epsilon);$$
$$C = \lim_{\epsilon\to 0+} C_\epsilon.$$

The operational meaning of C_ϵ is the maximum achievable rate at which one can communicate through a noisy channel with probability of error at most ϵ. Hence, C_ϵ and C can be alternatively defined as

$$C_\epsilon = \sup\{R : \forall\delta > 0, \exists n_0(\delta), \forall n \geq n_0(\delta), \exists(n, \exp(n(R-\delta)), \epsilon)\text{-code }\}$$

$$C = \sup\{R : \forall\epsilon > 0, \forall\delta > 0, \exists n_0(\delta,\epsilon), \forall n \geq n_0(\delta,\epsilon), \exists(n, \exp(n(R-\delta)), \epsilon)\text{-code }\}$$

Lower Bound

For a *Stationary Memoryless channel*,

$$C_\epsilon \geq \sup_{P_X} I(X;Y) \text{ for any } \epsilon \in (0,1]$$

Proof,
Fix any P_X on $\mathcal{A}$ and let $P_{X^n} = P_X^{\otimes n}$ be an IID product of the distribtuion P_X. From shannon's bound we have,

$$\epsilon \leq \mathbb{P}[i(X;Y) \leq \log M + \tau] + \exp(-\tau)$$

Wrt to the distribution $P_{X^n,Y^n} = P_{X,Y}^{\otimes n}$,

$$i\left(X^n;Y^n\right) = \sum_{k=1}^{n} i\left(X_k;Y_k\right),$$

The random variable $i(X^n, Y^n)$ is a sum of iids with mean $I(X;Y)$. By weak law of large numbers we have,

$$\mathbb{P}[i(X^n;Y^n) < n(I(X;Y)-\delta)] \to 0 \quad \text{for any } \delta > 0$$

Setting $\log M = n(I(X;Y)-2\delta)$ and $\tau = \delta n$ we get,

$$\epsilon_n \leq \mathbb{P}\left[\sum_{k=1}^{n} i\left(X_k;Y_k\right) \leq nI(X;Y)-\delta n\right] + \exp(-\delta n) \xrightarrow{n\to\infty} 0,$$

Hence, $\forall n$ s.t $\epsilon_n \leq \epsilon$,

$$\log M^*(n,\epsilon) \geq n(I(X;Y)-2\delta)$$

So,

$$C_\epsilon = \liminf_{n\to\infty} \frac{1}{n} \log M^*(n,\epsilon) \geq I(X;Y) - 2\delta$$

Since this holds for all P_X and $\delta > 0$, we can conclude $C_\epsilon \geq \sup I(X;Y)$