Quantum Information Theory I

Sibasish Ghosh

The Institute of Mathematical Sciences
CIT Campus, Taramani, Chennai 600 113, India.

Abstract

Classical information theory provides useful methods to process and transmit informations using classical physics. We will try to see through this set of three lectures how quantum theory helps us to have much more efficient ways to do these jobs of storing, processing and transmitting informations.

- **First lecture: Classical information involving Shannon's source coding and noisy channel coding theorem.**
- **Second lecture: Quantum information involving Schumacher's data compression.**
- **Third lecture: Quantum information involving pure state distillation and Holevo's bound.**

What is information?

Statement 1: The sun will rise on the east today.

Statement 2: We may have rain fall tonight at Bhubaneswar.

- **Statement 1 does not add any useful information to our knowledge, it is a certain event.**
- **Statement 2 does add some useful information to our knowledge, as rain fall at Bhubaneswar is not a certain event.**

Information is ignorance

- **Thus the amount of information about an event is the amount of ignorance (or, uncertainty) about that event.**
- **Ignorance increases with increase of the inverse of the probability p of the event.**
- **Total amount of ignorance of two independent events is sum of the ignorances.**
- **So the amount of ignorance $I(p)$ of should be additive function of p.**

Shannon entropy

- **Using continuity and additivity properties of $I(p)$, Shannon (1948) has shown that $I(p) = \log_2(1/p)$ upto some additive and/or multiplicative constant.**

- **So the average information content of a set X of n mutually exclusive but exaustive events x_1, x_2, , x_n with respective probabilities p_1, p_2, . . ., p_n is the given by the Shannon entropy $H(X) = \sum_{i=1}^{n} p_i \log_2 p_i$.**

- **$H(X)$ depends only on p_i's, not on event names x_i's.**

Certain event vs. most disordered event

- **For any random variable X with value set $\{x_1, x_2, \ldots, x_n\}$ and $\mathrm{Prob}(X = x_i) = p_i$, if x_i is a certain event then $H(X) = 0$; so we have no ignorance about X!**
- **For equally probable events x_1, x_2, ..., x_n, we have $H(X) = \log_2 n$; so we have maximum ignorance about X.**
- **For all other probability distributions, $0 \leq H(X) \leq \log_2 n$.**

String of random variables

- **The height h_i of a flight at time t_i, while moving from one place A to another place B, will not be far apart from its height at time $t_{i-1)}$ if $t_i - t_{i-1}$ is small enough. But during an entire year, the height X_i may vary within the interval $[h_i^{min}, h_i^{max}]$ for each i with associated probability $\mathrm{Prob}(X_i = h_i) \equiv p_i(h_i)$. So the random variables X_i, X_2, $\ldots$ are not independent. We need to know the joint probabilities $\mathrm{Prob}(X_1 = h_1, X_2 = h_2, \ldots, X_n = h_n)$ to get the information content about the heights of the flight.**

i.i.d. case

- **Classical information $\equiv$ Information content of some random variable** $X = \{X = x \text{ with } \mathrm{Prob}(X = x) = p(x)\}_x$**.**

- **For** L **i.i.d. random variables**
$X = \{X = x_i, p(x_i) | i = 1, 2, \ldots, m\}$**,** $H(X^L) = \sum_{i_1, i_2, \ldots, i_L = 1}^{m} p(x_{i_1}) p(x_{i_2}) \ldots p(x_{i_L}) \log_2(p(x_{i_1}) p(x_{i_2}) \ldots p(x_{i_L}))$**.**

Letters, alphabet, messages

- $X_1, X_2, \ldots, X_L$ **be i.i.d. random variables distributed as** X**. For any random variable** X**, each of its values** $x_1, x_2, \ldots, x_n$ **is called a letter; the set** $\{x_1, x_2, \ldots, x_n\}$ **is called alphabet; any string** $x_{i_1} x_{i_2} \ldots x_{i_L}$ **of length** L **is called a message where** L **can be any positive integer.** $\mathrm{Prob}(X^L = x_{i_1} x_{i_2} \ldots x_{i_L}) = p(x_{i_1}) p(x_{i_2}) \ldots p(x_{i_L})$**.**

- **Classical information theory deals with processing and transmitting messages (e.g., English language).**

Redundancy plays the role

- **Source coding:** How much a message can be compressed, i.e., how much redundancy is there in a message?
- **Channel coding:** How much redundancy one has to add to send any message through a noisy channel so that the receiver can decode the message reliably?

Variable vs. fixed length coding

- **Eight 'letters' 1, 2, 3, ..., 8 are produced by a source with respective probabilities 1/2, 1/4, 1/8, 1/16, 1/64, 1/64, 1/64, 1/64.**

Fixed lenght coding: $C(1) = 000$, $C(2) = 001$, $C(3) = 010$, $C(4) = 011$, $C(5) = 100$, $C(6) = 101$, $C(7) = 110$, $C(8) = 111$. **Average no. of bits per letter is 3.**

Variable lenght coding: $C(1) = 0$, $C(2) = 10$, $C(3) = 110$, $C(4) = 1110$, $C(5) = 111100$, $C(6) = 111101$, $C(7) = 111110$, $C(8) = 111111$. **Average no. of bits per letter is** $2 = H(X)$.

Classical data compression

- **For a binary source** $X:\ \mathrm{Prob}(X=0)=p,\ \mathrm{Prob}(X=1)=1-p$**. Any message is some** n**-bit string** $x_1x_2\ldots x_n$ **with** $x_i\in\{0,1\}$**.**
- **For large** n**, in a typical message, there will be** np **no. of** 0**'s. Total no. of such messages:** ${}^nC_{np}$**.**
- **Stirling approximation:** $\log_2{}^nC_{np}\approx nH(p,1-p)$**. So total no. of typical messages:** $2^{nH(p,1-p)}$**.**

Block coding

- **Each typical message (a block of n letters) can be encoded as a bit string of length $nH(p, 1-p)$.**
- **For large n, each typical message of length n occurs with probability $p^{np}(1-p)^{n(1-p)}$. So the total probability of occurrance of any one of the $2^{nH(p,1-p)}$ typical messages ≈ 1.**
- **We do not need to encode any atypical message, encoding the typical messages is enough!**

Block coding

- **Each typical message (a block of n letters) can be encoded as a bit string of length $nH(p, 1-p)$.**
- **For large n, each typical message of length n occurs with probability $p^{np}(1-p)^{n(1-p)}$. So the total probability of occurrance of any one of the $2^{nH(p,1-p)}$ typical messages ≈ 1.**
- **We do not need to encode any atypical message, encoding the typical messages is enough!**
- **$H(p, 1-p)$: no. of bits per letter required to express an arbitrarily large message, on an average.**

Asymptotic equipartition property

- $X_1, X_2, \ldots, X_n$ **are i.i.d. random variables according to** $X \equiv \{\mathrm{Prob}(X = x_i) = p(x_i) | i = 1, 2, \ldots, N\}$ **with** $E(X) < \infty$ **and** $E(X^2) < \infty$**.**
- **The random variable** Y_j **with values** $-\log_2 p(x_j)$ **will satisfy (by weak law of large numbers):** $(1/n)\sum_{j=1}^{n} Y_j$ **converges in probability to** $H(X)$**.**
- **Thus** $E((1/n)\sum_{j=1}^{n} Y_j) = -(1/n)\log_2 \mathrm{Prob}(X = x_1, X = x_2, \ldots, X = x_n) \rightarrow H(X)$ **as** $n \rightarrow \infty$**.**
- $x_1 x_2 \ldots x_n$ **is a typical message (or sequence).**

Typical sequences

- **Given ϵ, δ, for sufficiently large n, $x_1 x_2 \ldots x_n$ is a typical sequence if (i) $\mathrm{Prob}(x_1 x_2 \ldots x_n)$ satisfies $H(X) - \delta < -(1/n) \log_2 \mathrm{Prob}(x_1 x_2 \ldots x_n) < H(X) + \delta$ and (ii) the total probability of typical sequences exceeds $1 - \epsilon$.**
- **The random variable Y_j with values $-\log_2 p(x_j)$ will satisfy (by weak law of large numbers): $(1/n) \sum_{j=1}^{n} Y_j$ converges in probability to $H(X)$.**
- **Thus $E((1/n) \sum_{j=1}^{n} Y_j) = -(1/n) \log_2 \mathrm{Prob}(X = x_1, X = x_2, \ldots, X = x_n) \to H(X)$ as $n \to \infty$.**

Typical sequences (continued)

- **Total no. $N(\epsilon, \delta; n)$ of all such typical sequences satisfy $2^{n(H(X)+\delta)} \geq N(\epsilon, \delta; n) \geq (1-\epsilon)2^{n(H(X)-\delta)}$.**
- **Sum of the probabilities of all such typical sequences will lie between $1-\epsilon$ and 1.**
- **So, by using a block of length $n(H(X)+\delta)$ bits, we can encode all the typical sequences each of length n.**
- **No matter how atypical sequences are encoded, their total probability will be less than ϵ.**

Transmission through noisy channel

- To communicate messages over a noisy channel, we can improve reliability of transmission through redundancy (e.g., each bit may be sent many times so that the receiver can decode the message via majority vote.
- Given a channel, is it always possible to find a code which would ensure arbitrary reliability as the length of the message $\rightarrow \infty$)? What can be the rate (i.e., no. of bits required to encode each letter of the message) of such a code?

Shannon's noisy channel coding theorem

- **Shannon showed that any channel can be used for arbitrarily reliable communication at a non-zero rate provided there is some non-zero correlation between the input and the output.**
- **Shannon has also found a useful expression for the optimal rate that can be attained. This optimal rate is called the channel capacity.**

Binary symmetric channel

- $X = 0$ **and** 1 **with equal probability (taken for simplicity).**

$\mathrm{Prob}(\mathrm{Y} = 0|\mathrm{X} = 0) = \mathrm{Prob}(\mathrm{Y} = 1|\mathrm{X} = 1) = 1 - \mathrm{p}$ **and** $\mathrm{Prob}(Y = 0|X = 1) = \mathrm{Prob}(Y = 1|X = 0) = p$**.**

- **Here** $\mathrm{Prob}(Y = 0) = \mathrm{Prob}(Y = 0|X = 0)\ \mathrm{Prob}(X = 0) + \mathrm{Prob}(Y = 0|X = 1)\ \mathrm{Prob}(X = 1) = 1/2$**,** $\mathrm{Prob}(Y = 1) = \mathrm{Prob}(Y = 1|X = 0)\ \mathrm{Prob}(X = 0) + \mathrm{Prob}(Y = 1|X = 1)\ \mathrm{Prob}(X = 1) = 1/2$**. Using Bayes' rule:** $\mathrm{Prob}(X = 0|Y = 0) = (\mathrm{Prob}(Y = 0|X = 0)\ \mathrm{Prob}(X = 0))/\ \mathrm{Prob}(Y = 0) = 1 - p$**,** $\mathrm{Prob}(X = 0|Y = 1) = (\mathrm{Prob}(Y = 1|X = 0)\ \mathrm{Prob}(X = 0))/\ \mathrm{Prob}(Y = 1) = p$**,** $\mathrm{Prob}(X = 1|Y = 0) = (\mathrm{Prob}(Y = 0|X = 1)\ \mathrm{Prob}(X = 1))/\ \mathrm{Prob}(Y = 0) = p$**,** $\mathrm{Prob}(X = 1|Y = 1) = (\mathrm{Prob}(Y = 1|X = 1)\ \mathrm{Prob}(X = 1))/\ \mathrm{Prob}(Y = 1) = 1 - p$**.**

Binary symmetric channel (continued)

- **This gives** $H(X|Y) \equiv \text{Prob}(Y=0)H(X|Y=0) + \text{Prob}(Y=1)H(X|Y=1) = H(p,1-p)$**. Also we have** $H(X)=1$**. So** $I(X;Y) \equiv H(X) - H(X|Y) = 1 - H(p,1-p)$**.**
- **We want to construct a family of codes of increasing block size** n **such that the probability of decoding error goes to zero as** $n \to \infty$**.**
- **Assume that in this block coding, a** k**-bit string message is encoded as an** n**-bit string code word. So the total no. of code words** $= 2^k$ **and the rate** $R = k/n$**.**

Binary symmetric channel (continued)

- **The block coding must be such that the code words (after channel action) must be 'far apart'.**
- **For any n-bit code word to the channel, typically np of these bits will get corrupted (for large n), and their total no. is ${}^nC_{np} \approx 2^{nH(p,1-p)}$.**
- **For reliable decoding with rate R, $2^{nH(p,1-p)}.2^{nR} \leq 2^n$. So $R \leq I(X;Y) = 1 - H(p,1-p)$.**
- **The maximum rate R by which messages of large length can be sent throgh a noisy channel $\mathcal{N}$, with vanishing error of decoding, is the capacity $C(\mathcal{N})$ of the channel. For the binary symmetric channel $\mathcal{N}$, one can show that $C(\mathcal{N}) = 1 - H(p,1-p)$.**

Channel capacity for a general channel

- **For encoding of blocks of large size n, formed by letters x_1, x_2, ..., x_n of an alphabet $\mathcal{A}$ (where the corresponding random variable is $X = \{X = x,\ \mathrm{Prob}(X = x) = p_x | x \in \mathcal{A}\}$), and thereby sending these blocks through a noisy channel $\mathcal{N}$ (assuming, as in the case of binary symmetric channel, that the channel acts independently on individual letters – $\mathcal{N}$ is a 'memoryless' channel), one can show that the rate $R \leq \ \max\{I(X;Y) | X \text{ is input variable}\} \equiv C(\mathcal{N})$, where Y is the output of the channel $\mathcal{N}$.**

Channel capacity for a general channel (continued)

- **Among the $|\mathcal{A}|^n$ no. of messages, each of large size n, the total no. of typical messages will be $2^{nH(X)}$. Thus, once the receiver gets a message in Y^n (after the method of encoding the message from $X^n \rightarrow$ sending the encoded message through the channel $\mathcal{N}$), the total no. of typical messages that could have been sent is about $2^{nH(X|Y)}$ in no. Now each of such $2^{nH(X|Y)}$ typical input messages, can be sent (via $\mathcal{N}$) into an error sphere containing 2^{nR} no. of messages from Y^n, R being the rate of encoding. So, for error-free decoding, we must have $2^{nR} \times 2^{nH(X|Y)} \leq 2^{nH(X)}$, i.e., $R \leq I(X;Y)$. Thus $R \leq \max\{I(X;Y)|X\} \equiv C(\mathcal{N})$.**

Channel capacity for a general channel (continued)

- **Using random encoding, Shannon has shown that for large block size n, one can achieve the rate $R = C(\mathcal{N})$ for all noisy channels $\mathcal{N}$ with vanishing probability of decoding error provided there is some non-zero correlation among the inputs and outputs of the channel.**
- **Note that the capacity $C(\mathcal{N})$ is understood to be the maximum amount of information that the receiver would obtain about the input by knowing the output of the channel $\mathcal{N}$, and so, $C(\mathcal{N})$ has to be equal to $\max\{I(X;Y)|x\}$.**