(Classical) Information Theory III: Noisy channel coding

Sibasish Ghosh

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

Abstract

What is the best possible way to send any message through a noisy channel so that output of the channel is very much closed to input message? Shannon quantified it as the capacity of the channel – higher the capacity, higher will be the closedness of the output message to the input one, on an average. For this purpose, one will have to use the channel many times. Capacity of a channel is given by the maximum of the mutual information between the input and the output of the channel, over all possible input probability distributions. By using random block coding of sufficiently large length, one can always send messages reliably through the noisy channel at any rate which is less than or equal to the capacity of the channel. In the present lecture, we will discuss about the capacity of a noisy channel and how to achieve that.

Outline

- **What is a noisy channel?**
- **How does a channel work?**
- **Examples of noisy channels**
- **Essential idea to find out channel capacity**
- **Jointly typical sequences**
- **Noisy channel coding theorem and its converse**
- **Examples**
- **Summary**
- **References**

What is a noisy channel?

What is a channel: A channel is a device which gives an output if some input is fed into it. A telephone line connecting two (or more) telephone sets, broadcasting of programmes of a radio centre to some radio set, internet networking, nervous system, etc. are examples.

Noisy channel: A noisy channel is such a channel which, in general, distorts the input messages. While talking over a mobile phone inside a running electric train, we generally can't hear properly the voice of the other person, we are talking to. This happens due to the interference of the mobile phone network with the electromagnetic field caused by the overhead electric wire of the train. So, here the em field causes the noise in the mobile networking system.

How does a channel work?

(1) Express the message, to be sent through the channel, in terms of a string of letters chosen from an *apriori* fixed alphabet. (e.g., the sound wave of your speech over a telephone set gets transformed into a stream of several elementary electromagnetic signals) (2) This string of letters is then fed into the channel (The stream of em signals pass through the telephone cable). (3) Another string of letters is produced at the output of the channel (A stream of em signals is reproduced at the receiving end of the telephone line). (4) The output string of letters is then converted into a new message, which may or may not be same as the original message (The output stream of em signals are then converted into sound waves, which, in turn, may or may not be same as the input).

Examples of noisy channels

Example (1) Noiseless binary channel: Under the action of the channel, $0 \to 0$ and $1 \to 1$. So each bit can be sent through the channel in an error-free manner.

Example (2) Channel with non-overlapping outputs: Under the action of the channel, $0 \to 1$ with probability $1/2$, $0 \to 2$ with probability $1/2$, and $1 \to 3$ with probability $1/2$, $1 \to 4$ with probability $1/2$. Although the correspondence between the inputs and the outputs are not in one-to-one fashion, looking at the output, one can uniquely predict the input. So each of the bits can be sent through the channel in a noise-free manner.

Examples of noisy ....

Example (3) **Noisy English typing machine: Under the action of the channel, the letters are being transformed as: with probability** $1/2$**,** $a \to a$ **while with probability** $1/2$**,** $a \to b$**; with probability** $1/2$**,** $b \to b$ **while with probability** $1/2$**,** $b \to c$**;** $\ldots\ldots$**; with probability** $1/2$**,** $y \to y$ **while with probability** $1/2$**,** $y \to z$**; with probability** $1/2$**,** $z \to z$ **while with probability** $1/2$**,** $z \to a$**. But the punctuation symbols as well as the numbers remain intact. If we consider any paragraph involving only the letters (together with the punctuation symbols and numbers)** $a, c, \ldots, y$**, then the paragraph can be typed intact by the machine. So, in this case the channel will work in a noise-free manner on the letters** $a, c, \ldots, y$**.**

Examples of noisy ....

Example (4) **Binary symmetric channel: Given any proper fraction p, under the action of the channel, $0 \rightarrow 0$ with probability $(1-p)$, $0 \rightarrow 1$ with probability p, and symmetrically, $1 \rightarrow 1$ with probability $(1-p)$, $1 \rightarrow 0$ with probability p. Looking at the output, it is impossible to predict with certainty regarding the input. So, the bits can not be sent through this channel in an error-free manner.**

Example (5) **Binary erasure channel: Given any proper fraction p, under the action of the channel, $0 \rightarrow 0$ with probability $(1-p)$, 0 gets vanished with probability p, and $1 \rightarrow 1$ with probability $(1-p)$, 1 gets vanished with probability p. Looking at the vanished output, it is impossible to predict with certainty regarding the input.**

Examples of noisy ....

Example (6) Symmetric channel: Given any discrete probability distribution $\{p_1, p_2, \ldots, p_d\}$ **(thus** $p_1 + p_2 + \ldots + p_d = 1$ **and** p_i**'s are probabilities), under the action of the channel,** $1 \to 1$ **with probability** p_1**,** $1 \to 2$ **with probability** p_2**, ...,** $1 \to d$ **with probability** p_d**;** $2 \to 1$ **with probability** $p_1^{(2)}$**,** $2 \to 2$ **with probability** $p_2^{(2)}$**, ...,** $2 \to d$ **with probability** $p_d^{(2)}$**; ...;** $d \to 1$ **with probability** $p_1^{(d)}$**,** $d \to 2$ **with probability** $p_2^{(d)}$**, ...,** $d \to d$ **with probability** $p_d^{(d)}$**, where** $(p_1^{(2)}, p_2^{(2)}, \ldots, p_d^{(2)}), \ldots, (p_1^{(d)}, p_2^{(d)}, \ldots, p_d^{(d)})$ **are permutations of the row** $(p_1, p_2, \ldots, p_d)$ **while** $(p_2, p_2^{(2)}, \ldots, p_2^{(d)}), \ldots, (p_d, p_d^{(2)}, \ldots, p_d^{(d)})$ **are permutations of the column** $(p_1, p_1^{(2)}, \ldots, p_1^{(d)})$**. Example (4) is a special case of it.**

Essential idea to find out channel capacity

- **A channel is uniquely described by the entire set of transition probabilities $\mathrm{Prob}(Y = y|X = x) \equiv p_{y|x}$ for all possible values x of the input variable X and all possible values y of the output variable Y.**

Essential idea to find ....

- **Probing the output values y, we would like to know about the corresponding input values x. So, more we can reduce the content of information in X by knowing Y, more efficiently we can know X. Thus, we should look at the mutual information $I(X;Y)$. As, given any channel, we can alter only the input probability distributions, therefore, in order to get the maximum possible efficiency of the channel, we should maximize $I(X;Y)$ over all possible input distributions p_x's. This is maximized $I(X;Y)$ is the capacity of the channel in question.**

Essential idea to find ....

- **Choosing the appropriate input prob. distribution p_x is somewhat similar to choosing the selective no. of letters $a, c, \ldots, y$ in noisy English typing machine, so that looking at the perfectly distinguishable outputs, one can reliably infer about the inputs.**

Essential idea to find ....

- **The basic idea is to first encode each message i from the message set $\{1, 2, \ldots, M\}$ using some encoding procedure $X^n(i)$ to produce (in an 1-to-1 way) a string of n bits. Then send this n-bit string through the channel to produce another n-bit string Y^n. If $X^n(i)$ is a typical sequence, there are approximately $2^{nH(Y|X)}$ no. of possible sequences Y^n (forming the set $\mathcal{S}_{typical}$, say), each being of equal probability. As we want that no two X sequences produce the same Y sequence, and as, in total, there are $2^{nH(Y)}$ no. of typical Y sequences, the total no. of required disjoint sets $\mathcal{S}_{typical}$ must be $\leq 2^{n(H(Y)-H(Y|X))} = 2^{nI(X;Y)}$. So, we can send at most $2^{nI(X;Y)}$ no. of distinguishable X sequences of length n, for sufficiently large n.**

Jointly typical sequences

- **Discrete channel: Represented by $(\mathcal{X}, p_{y|x}, \mathcal{Y})$, with finite input as well as output sets $\mathcal{X}$, $\mathcal{Y}$ and transition probabilities $p_{y|x}$ such that $\sum_y p_{y|x} = 1$.**
- **n-th power discrete memoryless channel: Represented by $(\mathcal{X}^n, p_{y_1,y_2,\ldots,y_n|x_1,x_2,\ldots,x_n}, \mathcal{Y}^n)$ where $p_{y_k|x_1,x_2,\ldots,x_k;y_1,y_2,\ldots,y_{k-1}} = p_{y_k|x_k}$ (for $k = 1, 2, \ldots, n$), with x_j, y_j being the input, output of the j-th use of the channel.**
- **n-th power discrete channel without feedback: Represented by $(\mathcal{X}^n, p_{y_1,y_2,\ldots,y_n|x_1,x_2,\ldots,x_n}, \mathcal{Y}^n)$ where $p_{x_k|x_1,x_2,\ldots,x_{k-1};y_1,y_2,\ldots,y_{k-1}} = p_{x_k|x_1,x_2,\ldots,x_{k-1}}$ (for $k = 1, 2, \ldots, n$).**
- **So, for an n-th power DMC without feedback, we have: $p_{y_1,y_2,\ldots,y_n|x_1,x_2,\ldots,x_n} = \prod_{i=1}^{n} p_{y_i|x_i}$.**

Jointly typical ....

- **An (M, n) code for a discrete channel $(\mathcal{X}, p_{y|x}, \mathcal{Y})$ is represented by $(\{1, 2, \ldots, M\},\ X^n : \{1, 2, \ldots, M\} \rightarrow \mathcal{X}^n,\ g : \mathcal{Y}^n \rightarrow \{1, 2, \ldots, M\})$, X^n being the *encoding function* with $X^n(1), X^n(2), \ldots, X^n(M)$ being the *codewords* forming the *codebook* $\{X^n(1), X^n(2), \ldots, X^n(M)\}$ and g being the *decoding function*.**
- **Conditional probability of error for an (M, n) code for a discrete channel $(\mathcal{X}, p_{y|x}, \mathcal{Y})$: $\lambda_i \equiv \text{Prob}(g(Y^n) \neq i | X^n = X^n(i)) = \sum_{\{(y_1, y_2, \ldots, y_n) \in \mathcal{Y}^n : g((y_1, y_2, \ldots, y_n)) \neq i\}} p_{y_1, y_2, \ldots, y_n | x_1(i), x_2(i), \ldots, x_n(i)}$ for $i = 1, 2, \ldots, M$.**
- **Maximal probability of error for an (M, n) code for a discrete channel $(\mathcal{X}, p_{y|x}, \mathcal{Y})$:**
$\lambda^{(n)} \equiv \max \{\lambda_i : i = 1, 2, \ldots, M\}$.

Jointly typical ....

- **Average prob. of error for an (M, n) code:**

$P_{error}^{(n)} \equiv (1/M) \times \sum_{i=1} \lambda_i$**. Note that if all the indices i's are chosen uniformly from $\{1, 2, \ldots, M\}$, then** $P_{error}^{(n)} = \sum_{i=1}^{M} \text{Prob}(i \neq g(Y^n))$**. Obviously,** $P_{error}^{(n)} \leq \lambda^{(n)}$**. It can be shown that: smallness in P_{error} implies that in $\lambda^{(n)}$.**

- **Rate of an (M, n) code: $R \equiv (\log_2 M)/n$ bits per use of the channel.**

- **Achievable rate R: A given rate R is *achievable* if there exists a sequence of $(\lceil 2^{1R} \rceil, 1), (\lceil 2^{2R} \rceil, 2), \ldots, (\lceil 2^{nR} \rceil, n), \ldots$ codes for a discrete channel $(\mathcal{X}, p_{y|x}, \mathcal{Y})$ such that $\lambda^{(n)} \to 0$ as $n \to \infty$, where, for any real no. z, $\lceil z \rceil$ is the smallest integer $\geq z$.**

Jointly typical ....

- **(Operational) capacity C of a DMC without feedback: It is the supremum of all the achievable rates. Thus any rate $R < C$ will yield arbitrarily small probability of error for sufficiently large block lenght n.**

- **Intuitively, we decode the channel output Y^n as the index i if the codeword $X^n(i)$ is 'jointly typical' with Y^n.**

- **If $x^n \equiv (x_1, x_2, \ldots, x_n)$ is taken from the typical set $A^{(X)}(n, \epsilon)$ and the corresponding output sequence $y^n \equiv (y_1, y_2, \ldots, y_n)$ is also from the typical set $A^{(Y)}(n, \epsilon)$ for the joint prob. distribution p_{x^n, y^n}, (x^n, y^n) need not be jointly typical!**

Jointly typical ....

- **Jointly typical set:**

$A^{(X,Y)}(n,\epsilon) = \{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n \ : \ |-(1/n)\log_2 p_{x^n} - H(X)| < \epsilon, \ |-(1/n)\log_2 p_{y^n} - H(Y)| < \epsilon, \ |-(1/n)\log_2 p_{x^n,y^n} - H(X,Y)| < \epsilon\}$**, with** $p_{x^n,y^n} = \prod_{i=1}^{n} p_{x_i,y_i}$.

Jointly typical ....

- **Joint AEP: Let** (X^n, Y^n) **be drawn i.i.d. according to** $p_{x^n,y^n} = \prod_{i=1}^{n} p_{x_i,y_i}$**. Then (1)** $\text{Prob}(A^{(X,Y)}(n,\epsilon)) \to 1$ **as** $n \to \infty$**. (2)** $|A^{(X,Y)}(n,\epsilon)| \leq 2^{n(H(X,Y)+\epsilon)}$**. (3) If** $(\tilde{X}^n, \tilde{Y}^n)$ **has prob. distribution** $p_{x^n} \times p_{y^n}$**, then** $\text{Prob}((\tilde{X}^n, \tilde{Y}^n) \in A^{(X,Y)}(n,\epsilon)) \leq 2^{-n(I(X;Y)-3\epsilon)}$**. (4) For sufficiently large** n**,** $\text{Prob}((\tilde{X}^n, \tilde{Y}^n) \in A^{(X,Y)}(n,\epsilon)) \geq (1-\epsilon)2^{-n(I(X;Y)+3\epsilon)}$**.**

- **Using this AEP properties, a proof of Shannon's noisy channel coding theorem can be given. Shannon has shown in this proof that the 'operational capacity' of a noisy channel is asymptotically same as the 'informational capacity' of the channel.**

Noisy channel coding theorem and its converse

- **Given any rate R less than the capacity C, there exists a sequence $(\lceil 2^{nR} \rceil, n)$ of codes for a DMC without feedback with $\lambda^{(n)} \to 0$. Conversely, any sequence $(\lceil 2^{nR} \rceil, n)$ of codes for a DMC without feedback with $\lambda^{(n)} \to 0$, we must have $R \leq C$.**

Examples

- **(1) For the noiseless binary channel (example 1), each bit is transmitted undisturbed through the channel, and so, looking each time at the channel output, one will be able to infer about the channel input, without making any error. So, one can send exactly one bit undisturbed per single use of the channel. Therefore the capacity C of this channel must be 1 bit. This is easy to show by maximizing $I(X;Y)$ over all possible initial probability distributions p_x's. Choose, for example, $p_0 = 1/2, p_1 = 1/2$.**

Example ....

- **(2) For the noisy channel with non-overlapping outputs (example 2), each transmitted bit can be recovered, without making any error, by looking at each of the four outputs. So, one can send exactly one bit undisturbed per single use of the channel. Therefore the capacity C of this channel must be 1 bit. This is easy to show by maximizing $I(X;Y)$ over all possible initial probability distributions p_x's. Choose, for example, $p_0 = 1/2, p_1 = 1/2$.**

Example ....

- **(3) For the noisy English type writter (example 3), any set of all the alternate alphabets (there are 13 such alphabets are there) gets transmitted without making any error per single use of the channel. So, one can send exactly $\log_2 13$ bits of information intact per single use of the channel. Therefore the capacity C of this channel must be $\log_2 13$ bits. This is easy to show by maximizing $I(X;Y)$ over all possible initial probability distributions p_x's. Choose, for example, the uniform probability distribution: $p_a = p_b = \ldots = p_z = 1/(26)$.**

Example ....

- **(4) For the binary symmetric channel (example 4),** $I(X;Y) = H(Y) - H(Y|X) = H(Y) - \sum_{x=0,1} p_x H(Y|X = x) = H(Y) - \sum_{x=0,1} p_x \times H_2(p)$ **(as, for example,** $\text{Prob}(Y = 0|X = 0) = 1 - p$ **and** $\text{Prob}(Y = 1|X = 0) = p$**, therefore** $H(Y|X = 0) = H_2(p)$**)** $= H(Y) - H_2(p) \leq 1 - H_2(p)$ **(as** Y **is here a binary random variable, so** $H(Y) \leq \log_2 2 = 1$**), the inequality will become equality if** p_y **is uniform, which will hold good if and only if** p_x **is uniform,** *i.e.*, $p_0 = p_1 = 1/2$**. So** $C = 1 - H_2(p)$**.**

Summary

- **What do we mean by classical information of a data: it is given by the Shannon's entropy of the probability distribution of different events of that data.**
- **To what extent can one compress a data: for large string size of different events from the data, it is again given by the Shannon's entropy of the probability distribution of different events of that data, per unit length of the string.**
- **What is the maximum rate of transmission of data through a noisy channel: for sending any string of events of large length through the channel, the maximum rate can not cross the mutual information of the input and the output random variables per single use of the channel.**

References

- In this set of three lectures, I have mostly followed the book "Elements of Information Theory" by Thomas M. Cover and Joy A. Thomas (Wiley-India, 2006). This is an easily-readable book.

- There are few other good books on classical information theory: "Information Theory" by R. B. Ash (Interscience, New York, 1965). "Information Theory and Reliable Communication" by R. G. Gallager (John Wiley and Sons, Inc., New York, 1968). "Science and Information Theory" by L. Brillouin (Academic Press, New York, 1962).

References ....

- **One may also look at the freely downloadable book "Information Theory, Inference, and Learning Algorithms" by David J. C. Mackay (Cambridge Univ. Press, 2003) from the website address: http://www.inference.phy.cam.ac.uk/mackay/itila/book.html**