EE376A Information Theory Lecture 9 - 02/05/2015

Lecture 9: Communication and Channel Capacity

Lecturer: Tsachy Weissman *Scribe: Hanchel Cheng, Kyle Chiang, and Ashwin Siripurapu*

1 The Communication Problem

We model communication as the transmission of an input consisting of n symbols (denoted X^n) through a *noisy channel*. This channel will emit an output of n symbols (denoted Y^n). Think of the the channel as corrupting or adding noise to the input.

The behavior of the channel is characterized by the conditional probability distribution $P_{Y^n|X^n}$ which tells us the distribution over the outputs that it will emit for any given input.

Conceptually, we have the following picture:

$$X^n \longrightarrow \boxed{\begin{matrix}\text{noisy channel} \\ P_{Y^n|X^n}\end{matrix}} \longrightarrow Y^n$$

Our goal, then, is to find an encoder e and a decoder d such that we can take an input of m bits (called B^m), encode it using e to get an encoding of length n denoted X^n, pass the encoding through the channel to get Y^n, and then decode Y^n using d to recover an estimate of the original string of bits, denoted $\hat{B}^m$. Pictorially, that is:

$$(B_1, B_2, \ldots, B_m) = B^m \longrightarrow \boxed{\text{encoder } (e)} \xrightarrow{X^n} \boxed{\begin{matrix}\text{noisy channel} \\ P_{Y^n|X^n}\end{matrix}} \xrightarrow{Y^n} \boxed{\text{decoder } (d)} \longrightarrow \hat{B}^m = (\hat{B}_1, \hat{B}_2, \ldots, \hat{B}_m)$$

Ideally n will be small relative to m (we will not have to send a lot of symbols through the noisy channel), and the probability that the received message $\hat{B}^m$ does not match the original message B^m will be low. We will now make rigorous these intuitively good properties.

Definition 1. *We define a* scheme *to be a pair of encoder and decoder, denoted* (e, d).

Note that the definition of a scheme does not include the noisy channel itself. We must take the channel as it is given to us: we cannot modify it, but we can choose what symbols X^m to transmit through it.

Definition 2. *We define the* rate *of a scheme, denoted R, to be the number of bits communicated per use of the channel. This is equal to m/n in the notation of the diagram above.*

Definition 3. *We define the* probability of error *for a scheme, denoted P_e, to be the probability that the output of the decoder does not exactly match the input of the encoder. That is,*

$$P_e = P(B^m \neq \hat{B}^m).$$

Definition 4. *For a given channel, we say that a rate R is* achievable *if there exists a sequence of schemes $(e_1, d_1), (e_2, d_2), \ldots$, such that:*

1. *For all $n = 1, 2, \ldots$, scheme (e_n, d_n) has rate at least R, and*
2. $$\lim_{n\to\infty} P_e^{(n)} = 0,$$

   where $P_e^{(n)}$ denotes the probability of error of the n^{th} scheme.

2 Channel Capacity

We want to know what the best possible performance is under a particular noisy channel. This is essentially what channel capacity tells us.

Definition 5. *For a given channel, the* channel capacity *(denoted C) is the theoretical limit on the number of bits that can be reliably communicated (i.e., communicated with arbitrarily low probability of error) in one channel use.*

That is,

$$C \coloneqq \sup\{R : R \text{ is achievable } \}$$

We assume a "memoryless channel": $P_{Y^n|X^n}(y^n|x^n) = \prod_{i=1}^{n} P_{Y|X}(y_i|x_i)$ exhibiting a "single letter channel" characteristic of output symbol given input symbol. Restated, the i-th channel output only cares about the ith channel input.

$$X \sim P_x \longrightarrow \boxed{P_{Y|X}} \longrightarrow \text{random } Y$$

With this single letter channel, we now examine $I(X;Y)$. What distribution of X will maximize $I(X;Y)$ over all possible channel inputs?

$$C^{(I)} \triangleq \max_{P_X} I(X;Y)$$

$$\boxed{\text{Theorem: } C = C^{(I)}}$$

Example 6. Binary Symmetric Channel (BSC)

Let: $\mathcal{X} = \mathcal{Y} = \{0, 1\}$

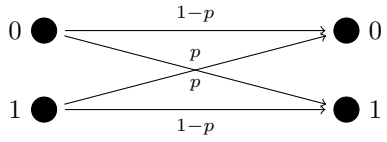
and the crossover probability be

$$P_{Y|X}(y|x) = \begin{cases} p & y \neq x \\ 1-p & y = x \end{cases}$$

$\iff$ channel matrix:

	0	1
0	$1-p$	p
1	p	$1-p$

$\iff$ bipartite graph:



$\iff Y = X \oplus_2 Z \leftarrow \text{ber(p)}$

To compute the capacity of the BSC, we first examine mutual information

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) && \text{definition of mutual information} \\ &= H(Y) - H(X \oplus_2 Z|X) && \text{substitute } Y \text{ with } X \oplus_2 Z \\ &= H(Y) - H(Z) && \text{given } X, X \oplus_2 Z \text{ is simply } Z \\ &= H(Y) - h_2(p) \leq 1 - h_2(p) && h_2(p) \text{ is binary entropy} \end{aligned}$$

To achieve equality, $H(Y) = 1$, i.e. Y is bernoulli $\frac{1}{2}$. Taking $X \sim Ber(\frac{1}{2})$ produces this desired Y and therefore gives $I(X;Y) = 1 - h_2(p)$

$$\implies C = 1 - h_2(p)$$

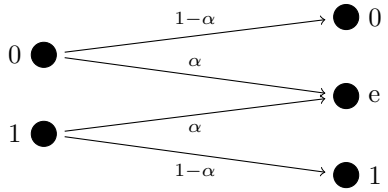
$(1 - h_2(p))n$ bits of information can be communicated reliably.

Example 7. Binary Symmetric Channel (BEC)

Let: $\mathcal{X} = \{0, 1\}, \mathcal{Y} = 0, 1, e$

and the crossover probability be

$$P_{Y|X}(y|x) = \begin{cases} \alpha & y = e \\ 1 - \alpha & y = x \end{cases}$$



$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(X) - [H(X|Y = e)P(Y = e) + H(X|Y = 0)P(Y = 0) + H(X|Y = 1)P(Y = 1)] \\ &= H(X) - [H(X\alpha + 0 + 0] \\ &= (1 - \alpha)H(x) \end{aligned}$$

To achieve equality, $H(X) = 1$, i.e. X is bernoulli $\frac{1}{2}$.

$$I(X;Y) = 1 - \alpha$$

$$\implies C = 1 - \alpha$$

$(1 - \alpha)n$ bits of information can be communicated reliably.

3 Extending Information Measures to Continuous R.V.

Definition 8. *Relative Entropy for two PDFs f and g, is defined as*

$$D(f||g) = \int f(x) \log \frac{f(x)}{g(x)} dx$$

Note: When integral on the right hand side is not well defined, $D(f||g) = \infty$

Definition 9. *Mutual Information between two continuous r.v. X, Y with joint pdf $f_{X,Y}$ is*

$$\begin{aligned} I(X;Y) &= D(f_{X,Y}||f_X \cdot f_Y) \\ &= \int \int f_{X,Y}(x,y) \log \frac{f_{X,Y}(x,y)}{f_X(x) f_Y(y)} dxdy \end{aligned}$$

3.1 Differential Entropy

Definition 10. *Differential entropy of X with PDF f_X is*

$$h(x) \triangleq E[-\log f_X(x)]$$

Example 11.

$$X \sim U[a,b]$$
$$f_X(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & otherwise \end{cases}$$
$$h(x) = \log(b-a)$$

Example 12.

$$X \sim N(0,\sigma^2), f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-x^2}{2\sigma^2}}$$

$$\begin{aligned} h(x) &= E\left[-\log\frac{1}{\sqrt{2\pi\sigma^2}} + \frac{x^2/2\sigma^2}{\ln 2}\right] \\ &= \frac{1}{2}\log 2\pi\sigma^2 + \frac{\sigma^2/2\sigma^2}{\ln 2} \\ &= \frac{1}{2}\log 2\pi\sigma^2 + \frac{1}{2\ln 2} \\ &= \frac{1}{2}\left[\log 2\pi\sigma^2 + \frac{\log e}{\log 2}\right] \\ &= \frac{1}{2}\left[\log 2\pi\sigma^2 + \log e\right] \\ &= \frac{1}{2}\log(2\pi\sigma^2 e) \end{aligned}$$

Note: differential entropies can be either positive or negative. The more correlated the random variable the more negative

Example 13. Significance of Differential entropy
Many Information theorists would argue none whatsoever. However, some others offer a different perspective. If you discretize $X \sim f$ into X_Δ with time period Δ,

$$P(x_\Delta = i) = \int_{i\Delta-\Delta/2}^{i\Delta+\Delta/2} f(x)dx \approx f(i\Delta)\cdot\Delta$$

$$\begin{aligned} H(X_\Delta) &= \sum_i P_i \log\frac{1}{P_1} \\ &\approx \sum_i f(i\Delta)\cdot\Delta\cdot\log\frac{1}{\Delta f(i\Delta)} \\ &= \log\frac{1}{\Delta} + \sum_i \left(f(i\Delta)\log\frac{1}{f(i\Delta)}\right)\Delta \\ H(X_\Delta) - \log\frac{1}{\Delta} &= \sum_i f(i\Delta)\log\frac{1}{f(i\Delta)}\Delta \xrightarrow{\Delta\to 0} \int f(x)\log\frac{1}{f(x)}dx = h(x) \end{aligned}$$

$$\implies H(X_\Delta) \approx \log\frac{1}{\Delta} + h(x)$$

Majority of the entropy for discretized system is accounted for with $\log\frac{1}{\Delta}$. The rest of it is $h(x)$ the differential entropy

References

[1] D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, UK, 2003.

[2] S. Geman and D. Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 6, pp. 721–741, 1984.