EE/Stats 376A: Information Theory **Winter 2017**

Lecture 14 — February 28

Lecturer: David Tse *Scribe: Sagnik M, Vivek B*

14.1 Outline

- Gaussian channel and capacity
- Information measures for continuous random variables

14.2 Recap

So far, we have focused only on communication channels with a discrete alphabet. For instance, the binary erasure channel (BEC) is a good model for links and routes in a network where packets of data are transferred correctly or lost entirely. The binary symmetric channel (BSC) on the other hand is quite an oversimplification of errors in a physical channel and therefore, not a very realistic model.

14.3 Gaussian Channel

In reality, a communication channel is analog, motivating the study of continuous alphabet communication channels. The physical layer of communication channels has additive noise due to a variety of reasons. Now, by the central limit theorem, the cumulative effect of a large number of small random effects is approximately normal, so modelling the channel noise as a Gaussian random variable is quite natural and valid in a large number of situations.

Mathematically, the Gaussian channel is a continuous alphabet time-discrete channel where the output Y_i is the sum of the input X_i and independent noise Z_i, where $Z_i \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$.

$$Y_i = X_i + Z_i.$$

14.3.1 Power Constraint

If we analyze the Gaussian channel like a discrete alphabet channel, the natural approach would be to try to find its capacity. However, if the input is unconstrained, we can choose an infinite subset of inputs arbitrarily far apart, so that they are distinguishable at the output with arbitrarily small probability of error. Such a scheme would then have infinite capacity.

Here we will impose an average power constraint on the input X. Any codeword $(X_1, \cdots, X_n)$ transmitted over the channel must satisfy

$$\mathbb{E}\big[\sum_{i=1}^{n} X_i^2\big] \leq nP. \tag{14.1}$$

14.3.2 Capacity

Theorem 1. In a Gaussian channel with average power constrained by P and noise distributed as $\mathcal{N}(0, \sigma^2)$, the channel capacity C is given by $\frac{1}{2}\log\left(1 + \frac{P}{\sigma^2}\right)$.

Theorem 1 is a cornerstone result for information theory, and we will devote this section to analyze the sphere packing structure of a Gaussian channel and provide an intuitive explanation of the theorem above. In the next lecture we will prove the theorem rigorously.

The sphere packing argument of the Gaussian channel is characterized by the noise and the output spheres. In particular, the ratio of the volume of the output sphere to the volume of a 'average' noise sphere gives us an upper bound on the codebook size, 2^{nR}.

Noise Spheres: Consider any input codeword X of length n and the received vector Y, where $Y_i = X_i + Z_i$, $Z_i \sim \mathcal{N}(0, \sigma^2)$. Then, the radius of the noise sphere is the distance between input vector X and received vector Y, equal to $\sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2} = \sqrt{\sum_{i=1}^{n} Z_i^2}$.

Now, $E[Z_i^2] = \text{Var}(Z_i) + E[Z_i]^2 = \sigma^2$ and since Z_i's are i.i.d by the Weak Law of Large Numbers, $\sum_{i=1}^{n} Z_i^2 \approx n\sigma^2$. Thus the 'average' radius of a noise sphere is approximately $\sqrt{n\sigma^2}$.

Output Sphere: Because of the power constraint P, the input sphere has radius at most $\sqrt{nP}$. The noise expands the input sphere into the output sphere by $n\sigma^2$. Since the output vectors have energy no greater than $nP + n\sigma^2$, they lie in a sphere of radius $\sqrt{n(P + \sigma^2)}$.

The volume of an n-dimensional sphere is given $C_n R^n$, where R is its radius and C_n is some constant. Therefore, we have

$$\begin{aligned}\text{Volume of Output Sphere} &= C_n(n(P+\sigma^2))^{\frac{n}{2}} \\ \text{Volume of Noise Sphere} &= C_n(n\sigma^2)^{\frac{n}{2}}.\end{aligned}$$

Therefore, the number of non-intersecting noise spheres in the output sphere is at most

$$\frac{C_n(n(P+\sigma^2))^{\frac{n}{2}}}{C_n(n\sigma^2)^{\frac{n}{2}}} = \left(1 + \frac{P}{\sigma^2}\right)^{\frac{n}{2}}.$$

For decoding with a low probability of error,

$$2^{nR} \leq \left(1 + \frac{P}{\sigma^2}\right)^{\frac{n}{2}} \implies R \leq \frac{1}{2}\log\left(1 + \frac{P}{\sigma^2}\right).$$

Similar to the case of discrete alphabet, we expect the expression $\frac{1}{2}\log(1 + \frac{P}{\sigma^2})$ to be the solution of a mutual information maximization problem with power constraint (14.1):

$$\begin{aligned}&\max_{f_X} I(X;Y) \\ &s.t \;\; E[X^2] \leq P,\end{aligned} \tag{14.2}$$

where $I(X;Y)$ is some notion of mutual information between **continuous** random variables X and Y. A rigorous proof of the result necessitates the introduction of the corresponding information measures (mutual information, KL-divergence, entropy etc.) for continuous random variables, which will be further used to solve the optimization problem (14.2) in the next lecture .

Definition 1. The quantity $\frac{P}{\sigma^2}$ is the *signal to noise ratio.*

14.4 Information Measures for Continuous RVs

14.4.1 Mutual Information

For discrete random variables $X, Y \sim p$, the mutual information $I(X;Y)$ equals $E\left[\log \frac{p(X,Y)}{p(X)p(Y)}\right]$. The probability mass function for discrete RVs is akin to the density function for continuous ones. This motivates the following definition.

Definition 2. *Mutual information* for continuous random variables $X, Y \sim f$ is given by

$$I(X;Y) \triangleq E\left[\log \frac{f(X,Y)}{f(X)f(Y)}\right]$$

.

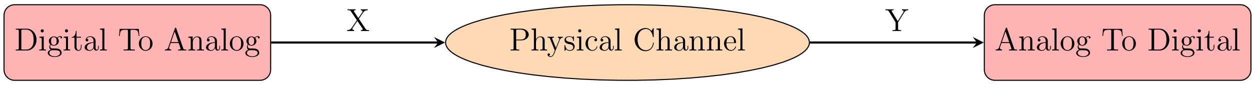

Figure 14.1: A typical flow-chart for a communication channel

Now, we will show that Definition 2 is sensible by proving that it's an approximation to the discretized form of X, Y. Since X and Y are the limit of arbitrarily small discretizations, we will have show that the definition is consistent with our previous definitions.

For $\Delta > 0$, define

$$\begin{aligned} X^\Delta &= i\Delta \quad \text{if} \quad i\Delta \le X < (i+1)\Delta, \\ Y^\Delta &= i\Delta \quad \text{if} \quad i\Delta \le Y < (i+1)\Delta. \end{aligned}$$

Now, for small Δ, we have $p(X^\Delta) \approx \Delta f(X), p(Y^\Delta) \approx \Delta f(Y)$ and $p(X^\Delta, Y^\Delta) \approx \Delta^2 f(X,Y)$, since f is the probability density function. Then,

$$\begin{aligned} I(X^\Delta, Y^\Delta) &= E\left[\log \frac{p(X^\Delta, Y^\Delta)}{p(X^\Delta)p(Y^\Delta)}\right] \\ &\approx E\left[\log \frac{f(X,Y)\Delta^2}{(f(X)\Delta)(f(Y)\Delta)}\right] \\ &= E\left[\log \frac{f(X,Y)}{f(X)f(Y)}\right] = I(X;Y) \\ \Longrightarrow \lim_{\Delta \to 0} I(X^\Delta, Y^\Delta) &= I(X;Y). \end{aligned}$$

From the above equations we see that Δ can be arbitrarily small, $I(X^\Delta, Y^\Delta)$ approximates $I(X;Y)$ to an arbitrary precision. Therefore,the definition for mutual information in continuous random variables is consistent with that for discrete ones.

14.4.2 Differential Entropy

The definition of mutual information immediately gives us

$$I(X;Y) = E\left[\log \frac{f(X,Y)}{f(X)f(Y)}\right] = E\left[\log \frac{f(Y|X)}{f(Y)}\right] = E\left[\log \frac{1}{f(Y)}\right] - E\left[\log \frac{1}{f(Y|X)}\right].$$

Akin to the discrete case, we would like to define the entropy of Y to be $E\left[-\log f(Y)\right]$ and the conditional entropy of Y given X to be $E\left[-\log f(Y|X)\right]$. However, the quantities $f(Y)$ and $f(Y|X)$ are not dimensionless. Furthermore, the entropy of a continuous random variable in its purest sense doesn't exist, since the number of random bits necessary to specify a real number to an arbitrary precision is infinite. It would however be very convenient to have a measure for continuous random variables, similar to entropy. This prompts the following definitions.

Definition 3. For a continuous random variable $Y \sim f$, the *differential entropy* is defined as

$$h(Y) \triangleq E\left[\log \frac{1}{f(Y)}\right]$$

Definition 4. For continuous random variables $X, Y \sim f$, the *differential conditional entropy* of Y conditioned on X is defined as

$$h(Y|X) \triangleq E\left[\log \frac{1}{f(Y|X)}\right]$$

Now, of course there are similarities and dissimilarities between entropy and differential entropy. We explore some of these in detail:

1. $H(X) \geq 0$ for any discrete random variable X. However, $h(X)$ need not be non-negative for every continuous random variable. This is because a probability mass function is always at most 1 but a density function can be arbitrarily large.

2. $H(X)$ is label-invariant. However, $h(X)$ need not be label invariant. The simplest change of labels, say $Y = aX$, for a scalar a proves this. Indeed the density function f_Y is given by $f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y}{a}\right)$ and so

$$h(Y) = E\left[\log \frac{1}{f_Y(Y)}\right] = E\left[\log \frac{|a|}{f_X(Y/a)}\right] = \log|a| + E\left[\log \frac{1}{f_X(X)}\right] = h(X) + \log|a|.$$

14.4.3 Differential Entropies of common distributions

We finish these lecture notes by computing the differential entropies of random variables with some common distributions.

1. Uniform Distribution. Suppose $X \sim \text{Unif}([0,a])$. Then, $f(x) = \frac{1}{a}$ for every $x \in [0,a]$, and $\log \frac{1}{f(x)} = \log a$. Then,

$$h(X) = E\left[\log \frac{1}{f(X)}\right] = E\left[\log \frac{1}{\frac{1}{a}}\right] = \log a.$$

2. Normal Distribution. Suppose $X \sim \mathcal{N}(0,1)$. Then, $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ for every x, and $\log_2 \frac{1}{f(x)} = \log_2 \sqrt{2\pi} + \frac{x^2}{2} \log_2 e$. The *differential entropy* of X is

$$h(X) = E\left[\log_2 \frac{1}{f(X)}\right] = E\left[\log_2 \sqrt{2\pi} + \frac{X^2}{2} \log_2 e\right]$$

$$\implies h(X) = \frac{1}{2}\log_2 2\pi + \frac{\log_2 e}{2}(\mathrm{Var}(X) + E[X]^2) = \frac{1}{2}\log_2 2\pi e.$$