Introduction to Information Theory

Lecture 7

Lecturer: Haim Permuter *Scribe: Yutzis Dov, Vortman Moti and Dor Tzur*

I. Differential Entropy and the Gaussian Channel

A. *Differential Entropy*

Let X be a random variable with a continuous alphabet.

- $F_X(x) = \Pr(X \leq x)$ - Cumulative Distribution Function {CDF}. $F(x)$ is a short notation for $F_X(x)$.
- $f_X(x) = \frac{dF_X(x)}{dx}$ - Probability Density Function {PDF} (in this course we will assume the derivative exists). $f(x)$ is a short notation for $f_X(x)$.

Definition 1 (Differential Entropy) The *differential entropy* $h(X)$ of a continuous random variable X with density $f_X(x)$ is defined as

$$h(X) \triangleq -\int f_X(x)\log_2(f_X(x))\,dx \triangleq \mathbb{E}[-\log_2 f_X]. \tag{1}$$

Question - can $h(X)$ be negative?

Example 1 (Uniform distribution) Let $X \sim U[0,a]$, i.e., $f_X(x) = \dfrac{1}{a}\cdot \mathbb{1}_{[0,a]}$

$$h(X) \;=\; -\int_0^a \frac{1}{a}\log_2 \frac{1}{a}dx = \log_2 a. \tag{2}$$

Remark 1 (Interpretation of entropy) For a finite alphabet r.v. X one can interpret entropy using the following result. The size (first order in the exponent) of the smallest set of sequences $\mathcal{A}^n$ such that $\lim_{n\to\infty}\Pr(\mathcal{A}^n) = 1$ is:

$$\lim_{n\to\infty} \log_2|\mathcal{A}^n| = nH(X). \tag{3}$$

For continuous alphabet a similar result hold but with volume of set instead of it's cardinality.

Definition 2 (Volume of the set) The *volume* of the set $A^n \subset \mathcal{R}^n$ is defined as:

$$\mathrm{Vol}\left(A^n\right) = \int_{X^n \in A^n} dx^n. \tag{4}$$

The volume (first order in the exp.) of the smallest set A^n such that $\lim_{n\to\infty} \Pr\left(A^n\right) = 1$ is:

$$\mathrm{Vol}\left(A^n\right) \approx 2^{nh(X)}. \tag{5}$$

This is rigorously stated and proved in Theorem 2. Using the outcome of example (1) we can show that Equation (5) holds:

- A^n is a cube of side length a, and of dimension n
- $\mathrm{Vol}\left(A^n\right) = a^n$

$$2^{nh(X)} = 2^{n \log_2 a} = a^n = \mathrm{Vol}\left(A^n\right).$$

Example 2 (Normal distribution) Find the differential entropy of $X \sim N\left(0, \sigma^2\right)$, i.e.,

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}.$$

Answer:

$$\begin{aligned}
h(X) &= -\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\dfrac{x^2}{2\sigma^2}} \left[\log_2 \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{x^2}{2\sigma^2} \log_2 e\right] dx \\
&= \frac{1}{2} \log_2 2\pi\sigma^2 + \frac{\sigma^2}{2\sigma^2} \log_2 e \\
&= \frac{1}{2} \log_2 2\pi e\sigma^2. \qquad (6)
\end{aligned}$$

Exercise 1 Let $\mathbf{X} \sim N\left(0, [K]\right)$, show that $h(X) = \frac{1}{2} \log_2 \left(2\pi e\right)^n |K|$ where n is the dimension of the square matrix $[K]$.

Definition 3 (Typical set) For $\epsilon > 0$ and any n, we define the *typical set* A_ϵ^n with respect to $f_X(x)$ as follows:

$$A_\epsilon^n = \left\{ X^n = (X_1, X_2, ..., X_n) \in \mathcal{X}^n : \left| -\frac{1}{n} \log_2 f(x^n) - h(X) \right| \leq \epsilon \right\}. \tag{7}$$

where $f(x^n) = \prod_{i=1}^n f_X\left(x_i\right)$.

The properties of the typical set for continuous random variables are similar to those for discrete random variables. The analogy of cardinality of typical set for the discrete case is the volume of the typical set for continuous random variable.

Theorem 1 (The typical set) The typical set A_ϵ^n has the following properties:

1) $\lim_{n\to\infty} \Pr\left(X^n \in A_\epsilon^n\right) = 1$.
2) $\mathrm{Vol}\left(A_\epsilon^n\right) \le 2^{n(h(X)+\epsilon)}$.
3) $\mathrm{Vol}\left(A_\epsilon^n\right) \ge (1-\epsilon)2^{n(h(X)-\epsilon)}$.

The proof is very similar to the finite alphabet case (see lecture 5) so here we will only prove (2) and the rest will be left for the reader.

Proof: (Continuous Alphabet)

$$\begin{aligned} 1 &= \int_{X^n} f(x^n)\, dx^n && (8)\\ &\overset{(a)}{\ge} \int_{x^n \in A_\epsilon^n} f(x^n)\, dx^n && (9)\\ &\overset{(b)}{\ge} \int_{x^n \in A_\epsilon^n} 2^{-n(h(X)+\epsilon)}\, dx^n && (10)\\ &= \mathrm{Vol}\left(A_\epsilon^n\right) 2^{-n(h(X)+\epsilon)} && (11)\end{aligned}$$

$$\Downarrow$$

$$\mathrm{Vol}\left(A_\epsilon^n\right) \le 2^{n(h(X)+\epsilon)}. \tag{12}$$

where

(a) follows from the fact that we are reducing the set.
(b) follows from the definition of A_ϵ^n.

■

Theorem 2 Let B_n be a set such that $\lim_{n\to\infty} \Pr\left(X^n \in B_n\right) = 1$, then for any $\eta > 0$

$$\frac{1}{n}\log_2 \mathrm{Vol}\left(B_n\right) \ge h(X) - \eta. \tag{13}$$

Proof: Let A_ϵ^n be a typical set, so we can claim that:

- $\Pr\left(X^n \in A_\epsilon^n\right) \to 1$ (Theorem 1)

- $\Pr\left(X^n \in B_n\right) \to 1$ (Assumption of theorem 2)

In other words $\forall \delta > 0, \exists n$ large enough such that:

$$\Pr\left(X^n \in A_\epsilon^n\right) > 1 - \delta \tag{14}$$

$$\Pr\left(X^n \in B_n\right) > 1 - \delta \tag{15}$$

$$\begin{aligned}
\Pr\left(X^n \in (B_n \cap A_\epsilon^n)\right) &= \Pr\left(X^n \in B_n\right) + \Pr\left(X^n \in A_\epsilon^n\right) - \Pr(X^n \in B_n \cup A_\epsilon^n) \\
&\overset{(a)}{\geq} \Pr\left(X^n \in B_n\right) + \Pr\left(X^n \in A_\epsilon^n\right) - 1 \\
&\overset{(b)}{\geq} 1 - \delta + 1 - \delta - 1 = 1 - 2\delta.
\end{aligned}$$

where

(a) follows from the fact that $\Pr(X^n \in B_n \cup A_\epsilon^n) \leq 1$.

(b) follows from Equation (14) and (15).

$$\begin{aligned}
1 - 2\delta &\leq \Pr\left(X^n \in (B_n \cap A_\epsilon^n)\right) \\
&= \int_{x^n \in (B_n \cap A_\epsilon^n)} f(x^n)\, dx^n \\
&\overset{(a)}{\leq} \int_{x^n \in (B_n \cap A_\epsilon^n)} 2^{-n(h(X)-\epsilon)}\, dx^n \\
&\overset{(b)}{\leq} \int_{x^n \in B_n} 2^{-n(h(X)-\epsilon)}\, dx^n
\end{aligned}$$

$$\Downarrow$$

$$\text{Vol}\left(B_n\right) \geq (1 - 2\delta) 2^{n(h(X)-\epsilon)}.$$

where

(a) follows from Theorem 1

(b) follows from the fact that we are increasing the volume.

We can choose the value of $\delta > 0$ and $\epsilon > 0$, as small as we like. Let $\epsilon = \eta$ and $\delta \to 0$ so that $(1 - 2\delta) \to 1$. Choosing those values of δ and ϵ will yield the desired result :

$$\text{Vol}\left(B_n\right) \geq 2^{n(h(X)-\eta)}. \tag{16}$$

■

Definition 4 (Divergence) Divergence between two PDFs $f(x)$ and $g(x)$ that satisfy that if for some x, $g(x) = 0$, then $f(x) = 0$ is defined as

$$D\left(f_X \| g_X\right) \triangleq \int_{x \in \mathcal{X}} f_X(x) \log_2 \frac{f_X(x)}{g_X(x)}\, dx. \tag{17}$$

Lemma 1 (Non-negativity of divergence) Divergence is non-negative:

1) $D\left(f_X \| g_X\right) \geq 0$.
2) $D\left(f_X \| g_X\right) = 0 \Leftrightarrow f_X(x) = g_X(x)$.

Proof:

$$\begin{aligned} -D\left(f_X \| g_X\right) &= \int_{x \in \mathcal{X}} f_X(x) \log_2 \frac{g_X(x)}{f_X(x)}\, dx \\ &= \mathbb{E}_f\left[\log_2 \frac{g_X}{f_X}\right] \overset{(a)}{\leq} \log_2 \mathbb{E}_f\left[\frac{g_X}{f_X}\right] = 0 \end{aligned} \tag{18}$$

where

(a) follows from Jensen's inequality.

We have equality iff we have equality in Jensen's inequality, which occurs iff $f_X(x) = g_X(x)$ almost everywhere (i.e. there is a countable set of x $\in \mathcal{X}$ for which $f_X(x) \neq g_X(x)$). ■

Definition 5 (Conditional Entropy)

$$h\left(X|Y\right) \triangleq -\int f_{X,Y}\left(x,y\right) \log_2 f_{X|Y}\left(x|y\right)\, dx\, dy = \mathbb{E}\left[-\log_2 f_{X|Y}\left(X|Y\right)\right]. \tag{19}$$

Definition 6 (Mutual Information) The $mutual\ information$ between two random variables X and Y is given by

$$I(X;Y) \triangleq D\left(f_{X,Y} \| f_X f_Y\right) = h\left(X\right) - h\left(X|Y\right). \tag{20}$$

Alternatively and equivalent

$$I\left(X;Y\right) = \sup_{\mathcal{Q},\mathcal{P}} I\left([X]_{\mathcal{P}};[Y]_{\mathcal{Q}}\right). \tag{21}$$

Where the supremum is over all finite partitions $\mathcal{P}$ and $\mathcal{Q}$.

The quantization of X by $\mathcal{P}$ (denoted $[X]_{\mathcal{P}}$) is the discrete random variable defined by

$$P\left(x_i\right) = \int_{x_i - \Delta}^{x_i + \Delta} f_X(x)\, dx. \tag{22}$$

Lemma 2 (Mutual Information is non-negative)

$$I(X;Y) \geq 0. \tag{23}$$

Or equivalently,

$$h(X) \geq h(X|Y). \tag{24}$$

Exercise 2 (Property of covariance matrix) Prove that the Determinant of a covariance matrix is less than or equal to the product of it's diagonal elements.

Answer: Using the inequality $h(X^n) \leq \sum_{i=1}^{n} h(X_i)$ where X^n is Gaussian random vector with covariance matrix K we obtain the following outcome :

$$\frac{n}{2}\log_2 2\pi e\, |K|^{\frac{1}{n}} \leq \sum_{i=1}^{n} \frac{1}{2}\log_2 2\pi e K_{ii}, \tag{25}$$

hence

$$|K| \leq \prod_{i=1}^{n} K_{ii}. \tag{26}$$

Lemma 3

$$h(aX) = h(X) + \log_2(|a|). \tag{27}$$

Proof: Let $Y = aX$. Then,

$$f_Y\,(y) = \frac{1}{|a|} f_X\left(\frac{y}{a}\right). \tag{28}$$

and

$$\begin{aligned} h(Y) &= -\int f_Y(y)\log_2 f_Y(y)\,dy \\ &= -\int \frac{1}{a} f_X\left(\frac{y}{a}\right)\log_2\left(\frac{1}{a} f_X\left(\frac{y}{a}\right)\right) a\, d\left(\frac{y}{a}\right) \\ &= -\int f_X\left(\frac{y}{a}\right)\log_2\left(f_X\left(\frac{y}{a}\right)\right)\, d\left(\frac{y}{a}\right) + \log_2(|a|) \\ &= h(X) + \log_2(|a|). \end{aligned} \tag{29}$$

■

Lemma 4 $h(A\mathbf{X}) = h(\mathbf{X}) + \log_2(|det(A)|)$ (Left as an exercise for the reader).[1]

Lemma 5 (Maximum entropy) Let $X \sim f_X(x)$ be a random variable with $E(X) = 0$ and $E(X^2) = \sigma^2$ then $h(X) \leq \frac{1}{2}\log_2 2\pi e\sigma^2$ and equality holds iff $X \sim N(0, \sigma^2)$

Proof:

Let $g(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$

$$\begin{aligned} 0 \leq D(f_X \| g_X) &= \int_{x\in\mathcal{S}} f_X(x) \log_2 \frac{f_X(x)}{g_X(x)}\, dx \\ &= \int_{x\in\mathcal{S}} f_X(x) \log_2 f_X(x)\, dx - \int_{x\in\mathcal{S}} f_X(x) [\log_2 \frac{1}{\sqrt{2\pi\sigma^2}} - \log_2 e^{\frac{x^2}{2\sigma^2}}]\, dx \\ &= -h(X) + \frac{1}{2}\log_2 2\pi\sigma^2 + \frac{1}{2}\log_2 e = -h(X) + \log_2 2\pi e\sigma^2 \end{aligned}$$

$$\Downarrow$$

$$h(X) \leq \tfrac{1}{2}\log_2 2\pi e\sigma^2. \tag{30}$$

■

Lemma 6 for any random variable X and estimator $\hat{X}$

$$\mathbb{E}[(X - \hat{X})^2] \geq \frac{2^{2h(X)}}{2\pi e}. \tag{31}$$

Proof: from last lemma we derive $\sigma^2 \geq \frac{2^{2h(X)}}{2\pi e}$ so

$$\mathbb{E}[(X - \hat{X})^2] \overset{(a)}{\geq} \mathbb{E}[(X - E(X))^2] = var(X). \tag{32}$$

(a) follows from the fact that $E[X]$ is the best estimator of X.

■

B. Gaussian Channel

The most important continuous alphabet channel is the Gaussian channel depicted in Figure (1). This is a discrete time channel with output Y_i at time i, where Y_i is the sum of the input X_i and the white noise Z_i. The noise Z_i is drawn i.i.d from a Gaussian distribution with variance σ_z^2. Thus,

[1] when we talk about h(x,y) we assume that f(x,y) exits

$$Y_i = X_i + Z_i, \quad Z_i \sim \mathcal{N}(0, \sigma_z^2). \tag{33}$$

The noise Z_i is assumed to be independent of the signal X_i. The most common limitation on the input is an energy or power constraint. We assume an average power constraint. For any codeword $x^n = (x_1, x_2, ..., x_n)$ transmitted over the channel, we require that

$$\frac{1}{n}\mathbb{E}[\sum_{i=1}^{n} x_i^2] \le P. \tag{34}$$

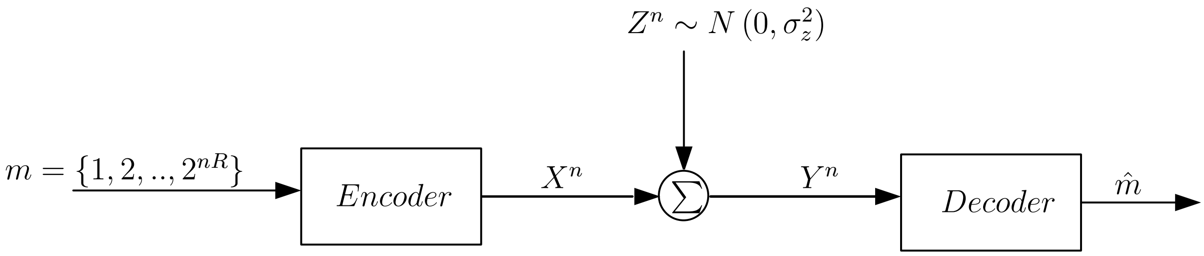

Fig. 1. Communication system with AWGN

- $Enc : \{1, 2, .., 2^{nR}\} \to \mathbb{R}^n$.
- $Dec : \mathbb{R}^n \to \{1, 2, .., 2^{nR}\}$.

We now define the (information) capacity of the channel as the maximum of the mutual information between the input and output over all distributions on the input that satisfy the power constraint.

Definition 7 (Achievable Rate) A rate R is said to be *achievable* if there exists a sequence of $(2^{nR}, n)$ codes such that $Pr(M \neq \hat{M}) \to 0$.

Definition 8 (Operational Capacity) Operational capacity C is the supremum over all achievable rates.

Definition 9 (Information Capacity) The information capacity of the Gaussian channel with power constraint P is

$$C = \max_{f(x):E[X^2]\leq P} I(X;Y). \tag{35}$$

Example 3 Let us compute C for the Gaussian case:

$$\begin{aligned}
I(X;Y) &= h(Y) - h(Y|X) \\
&= h(Y) - h(Y - X|X) \\
&= h(Y) - h(Z|X) \\
&= h(Y) - h(Z) \\
&\overset{(a)}{\leq} \frac{1}{2}\log_2 2\pi e(\sigma_z^2 + P) - \frac{1}{2}\log_2 2\pi e(\sigma_z^2) \\
&= \frac{1}{2}\log_2(\frac{\sigma_z^2 + P}{\sigma_z^2}) \\
&= \frac{1}{2}\log_2(1 + SNR).
\end{aligned} \tag{36}$$

where step (a) follows from the fact that X is independent of Z therefore $\mathbb{E}[Y^2] = \mathbb{E}[X^2] + \mathbb{E}[Z^2]$. In addition the differential entropy is bounded according to Lemma 5. Finally, note that the upper bound in (36) is achieved with equality if $X \sim \mathcal{N}(0, P)$.