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 that 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 $f(x)$ is defined as

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

(1)

Question - can $h(X)$ be lower than zero?

**Example 1 (Uniform distribution.)** Let $X \sim \text{Unif}[0, a]$, i.e., $f_X(x) = \frac{1}{a}$.

$$h(X) = -\int_0^a \frac{1}{a} \log_2 \frac{1}{a} \, dx = \log_2 a$$

(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 exp.) of the smallest sequence of set $A^n$ such that $\lim_{n \to \infty} \Pr(A^n) = 1$ is:

$$\lim_{n \to \infty} \log_2 |A^n| = n H(X)$$

(3)

For continuous alphabet a similar results hold but with volume of set instead of size of set.
Definition 2 (Volume of the set) The volume of the set $A^n \subset \mathbb{R}^n$ is defined as:

$$\text{Vol} (A^n) = \int_{X^n \in A^n} dx_1 dx_2 \ldots dx_n \quad (4)$$

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

$$\text{Vol} (A^n) \approx 2^{nh(X)} \quad (5)$$

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 size $a$, and $\text{dim} = n$
- $\text{Vol} (A^n) = a^n$

$$2^{nh(X)} = 2^n \log_2 a = a^n = \text{Vol} (A^n)$$

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

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

Answer:

$$h(X) = - \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{x^2}{2\sigma^2}} \left[ \log_2 \frac{1}{\sqrt{2\pi \sigma^2}} - \frac{x^2}{2\sigma^2} \log_2 e \right] \, dx $$

$$= \frac{1}{2} \log_2 2\pi \sigma^2 + \frac{\sigma^2}{2 \sigma^2} \log_2 e$$

$$= \frac{1}{2} \log_2 2\pi e \sigma^2 \quad (6)$$

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

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

$$A^n_\epsilon = \left\{ X^n = (x_1, x_2, \ldots, x_n) \in S^n : \left| -\frac{1}{n} \log_2 f(x^n) - h(x) \right| \leq \epsilon \right\} \quad (7)$$
where \( f(x^n) = \prod_{i=1}^{n} f_X(x_i) \).

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 \( A^n_\epsilon \) has the following properties:

1. \( \lim_{n \to \infty} \Pr(X^n \in A^n_\epsilon) = 1 \)
2. \( \text{Vol}(A^n_\epsilon) \leq 2^{nh(x)+\epsilon} \)
3. \( \text{Vol}(A^n_\epsilon) \geq (1-\epsilon)2^{nh(x)-\epsilon} \)

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

**Proof:** (Continuous Alphabet.)

\[
1 = \int_{X^n} f(x^n) \, dx^n
\]

(a)

\[
\geq \int_{X^n \in A^n_\epsilon} f(x^n) \, dx^n \quad (9)
\]

(b)

\[
= \text{Vol}(A^n_\epsilon) 2^{-n(h(X)+\epsilon)} \quad (10)
\]

\[
\downarrow
\]

\[
\text{Vol}(A^n_\epsilon) \leq 2^{nh(x)+\epsilon} \quad (12)
\]

where

(a) follows from the fact that we are reducing the set.

(b) follows from the definition of \( A^n_\epsilon \).

\[ \blacksquare \]

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

\[
\frac{1}{n} \log_2 \text{Vol}(B_n) \geq h(X) - \eta
\]

(13)
Proof: Let $A^n_\epsilon$ be a typical set, so we can claim that:

- $\Pr (X^n \in A^n_\epsilon) \to 1$ (Theorem 1)
- $\Pr (X^n \in B_n) \to 1$ (Assumption of theorem 2)

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

$$\Pr (X^n \in A^n_\epsilon) > 1 - \delta$$  \hspace{1cm} (14)

$$\Pr (X^n \in B_n) > 1 - \delta$$  \hspace{1cm} (15)

$$\Pr (X^n \in (B_n \cap A^n_\epsilon)) = \Pr (X^n \in B_n) + \Pr (X^n \in A^n_\epsilon) - \Pr(X^n \in B_n \cup A^n_\epsilon) \geq \Pr (X^n \in B_n) + \Pr (X^n \in A^n_\epsilon) - 1$$

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

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

$$1 - 2\delta \leq \Pr (X^n \in (B_n \cap A^n_\epsilon))$$

$$\leq \int_{X^n \in (B_n \cap A^n_\epsilon)} f(x^n) \, dx^n$$

$$\leq (a) \int_{X^n \in (B_n \cap A^n_\epsilon)} 2^{-n(h(X) - \epsilon)} \, dx^n$$

$$\leq (b) \int_{X^n \in B_n} 2^{-n(h(X) - \epsilon)} \, dx^n$$

$$\Downarrow$$

$$\operatorname{Vol} (B_n) \geq (1 - 2\delta) 2^{n(h(x) - \epsilon)}$$  \hspace{1cm} (16)

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 be $\epsilon = \eta$ and $\delta \to 0$ so that $(1 - 2\delta) \to 1$. Choosing those values of $\delta$ and $\epsilon$ will yield the desired result:

$$\operatorname{Vol} (B_n) \geq 2^{n(h(x) - \eta)}$$  \hspace{1cm} (16)
Definition 4 (Divergence) Divergence between two pdf’s $f(x)$ and $g(x)$ that satisfy that if for some $x$, $g(x) = 0$, then $f(x) = 0$ is defined as

$$D (f_X \| g_X) \triangleq \int_{x \in S} f_X(x) \log_2 \frac{f_X(x)}{g_X(x)} \, dx$$

(17)

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

1) $D (f_X \| g_X) \geq 0$
2) $D (f_X \| g_X) = 0 \iff f_X(x) = g_X(x)$

Proof:

$$-D (f_X \| g_X) = \int_{x \in S} f_X(x) \log_2 \frac{g_X(x)}{f_X(x)} \, dx$$

$$= \mathbb{E}_f \left[ \log_2 \frac{g_X}{f_X} \right] \overset{(a)}{\leq} \log_2 \mathbb{E}_f \left[ \frac{g_X}{f_X} \right] = 0$$

(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).

Definition 5 (Condition Entropy) The condition entropy $h(X \mid Y)$ between two random variables $X$ and $Y$ is given by

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

(19)

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

$$I(X; Y) \triangleq D [f_{X,Y} \| f_X f_Y] = h(X) - h(X \mid Y)$$

(20)

Alternatively and equivalent

$$I(X; Y) = \sup_{Q,P} I ([X]_P ; [Y]_Q)$$

(21)
Where the suprimum is over all finite partitions \( P \) and \( Q \).

The quantization of \( X \) by \( P \) (denoted \( [X]_P \)) is the discrete random variable defined by

\[
P(x_i) = \int_{x_i - \Delta}^{x_i + \Delta} f_X(x) \, dx
\]  

(22)

**Lemma 2 (Mutual Information is non-negative)**

\[
I(X;Y) \geq 0
\]  

(23)

Or equivalent

\[
h(X) \geq h(X|Y)
\]  

(24)

**Exercise 2 (Property of covariance matrix)** Proof that the Determinant of a covariance matrix is less or equal to the product of the diagonal elements

Ans: 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 :

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

(25)

hence

\[
|K| \leq \prod_{i=1}^{n} K_{ii}
\]  

(26)

**Lemma 3**

\[
h(aX) = h(X) + \log_2(|a|)
\]  

(27)

**Proof:** Let \( Y = aX \).

Then \( f_Y(y) = \frac{1}{|a|} f_X \left( \frac{y}{a} \right) \), and

\[
h(Y) = -\int f_Y(y) \log_2 f_Y(y) \, dy
\]

\[
= -\int f_X \left( \frac{y}{a} \right) \log_2 \left( \frac{1}{|a|} f_X \left( \frac{y}{a} \right) \right) \, \frac{a}{y} \, d \left( \frac{y}{a} \right)
\]

\[
= -\int f_X \left( \frac{y}{a} \right) \log_2 f_X \left( \frac{y}{a} \right) \, d \left( \frac{y}{a} \right) + \log_2(|a|)
\]

\[
= h(X) + \log_2(|a|)
\]  

(28)
Lemma 4 \( h(AX) = h(X) + \log_2(|\det(A)|) \) (Left as an exercise for the reader)\(^1\)

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

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

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

\[
\downarrow
h(X) \leq \frac{1}{2} \log_2 2\pi e \sigma^2 \tag{29}
\]

Lemma 6 \( \mathbb{E}[(X - \hat{X})^2] \geq \frac{2^{2h(X)}}{2\pi e} \)

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

\[
\mathbb{E}[(X - \hat{X})^2] \geq \mathbb{E}[(X - E(X))^2] = \text{var}(X) \tag{30}
\]

B. Gaussian Channel

The most important continuous alphabet channel is the Gaussian channel depicted in Figure (1). This is a time discrete channel whit 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 \( \sigma_z^2 \). Thus,

\(^1\)when we talk about \( h(x,y) \) we assume that \( f(x,y) \) exits
\[ Y_i = X_i + Z_i, \quad Z_i \sim \mathcal{N}(0, \sigma_z^2) \] 

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

\[ \frac{1}{n} \mathbb{E} \left[ \sum_{i=1}^{n} x_i^2 \right] \leq P \] 

Fig. 1. Communication system with WGN

- \( En: \{1, 2, ..., 2^{nR}\} \to \mathbb{R}^n \)
- \( De: \mathbb{R}^n \to \{1, 2, ..., 2^{nR}\} \)

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.)** \( R \) is an achievable rate if there exists a sequence of \((2^{nR}, n)\) codes such that \( P(M \neq \hat{M}) \to 0 \).

**Definition 8 (Operational Capacity.)** Operational capacity \( C \) is the suprimum over all achievable rates.
**Definition 9 (Information Capacity.)** The information capacity of the Gaussian channel with power constraint $P$ is

$$ C = \max_{f(x): E(X^2) \leq P} I(X; Y) \quad (33) $$

**Example 3** Let us compute $C$ for the Gaussian case:

$$ I(X; Y) = h(Y) - h(Y|X) $$$$ = h(Y) - h(Y - X|X) $$$$ = h(Y) - h(Z|X) $$$$ = h(Y) - h(Z) $$$$ \leq (a) \frac{1}{2} \log_2 2\pi e (\sigma_z^2 + P) - \frac{1}{2} \log_2 2\pi e (\sigma_z^2) $$$$ = \frac{1}{2} \log_2 \left( \frac{\sigma_z^2 + P}{\sigma_z^2} \right) $$$$ = \frac{1}{2} \log_2 (1 + SNR), \quad (34) $$

where step (a) follows from the fact that $X$ is independent of $Z$ therefore $E[Y^2] = E[X^2] + E[Z^2]$. In addition the differential entropy is bounded according to Lemma 5.

Finally, note that the upper bound in (34) is achieved with equality if $X \sim N(0, P)$.