LECTURE 16

Last time:

- Continuous Random Variables
- Differential Entropy
- Properties of differential entropy

Lecture outline

- More on Differential Entropy
- AEP for continuous random variables
- Coding Theorem
- Gaussian Channels

Review

- Differential entropy

$$h(X) = -\int f_X(x) \log f_X(x) dx$$

- Differential entropy does not give the absolute amount of randomness, but rather a relative measure.
- Differential entropy of a continuous r.v. depends on how the r.v. is represented.
- Properties of Differential entropy
  - Chain rule
  - Information Inequality
  - Conditioning reduces entropy
- For X taking value in $[a, b]$, uniform distribution maximizes the differential entropy.

Maximizing Entropy

For any r.v. X' taking values in $[a, b]$, let X be uniformly distributed,

$$h(X) - h(X') = D(X' \| X)$$

For a zero-mean r.v. X with $E(X^2) = \sigma^2$, what distribution maximizes the differential entropy?

$$\max_f \left[- \int f(x) \log f(x) dx \right]$$

subject to the constraint

$$\int f(x) dx = 1$$
$$\int x f(x) dx = 0$$
$$\int x^2 f(x) dx = \sigma^2$$

Can be solved by Lagrange method to conclude $X \sim N(0, \sigma^2)$.

Gaussian Random Variables

Gaussian distribution maximizes the differential entropy for the same first and second order moment.

Assume X' has the same first and second order moment as the Gaussian random variable X. Let the density of X be f and density of X' be g.

$$
\begin{aligned}
& D(X'||X) = D(g||f) \\
= & \int g(x) \log \frac{g(x)}{f(x)} dx \\
= & \int g(x) \log g(x) dx \\
& - \int g(x) \left[\log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}(x-\mu)^2 \right] dx \\
= & -h(X') - \int f(x) \left[\log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}(x-\mu)^2 \right] dx \\
= & h(X) - h(X')
\end{aligned}
$$

Jointly Gaussian Random Variables

Let $\underline{W}$ be a random vector with i.i.d. $N(0,1)$ entries.

$$h(\underline{W}) = \frac{1}{n}\log(2\pi e)^n$$

Let $\underline{X}$ be a Gaussian random vector with mean $\underline{\mu}$ and covariance matrix

$$E[(\underline{X}-\underline{\mu})(\underline{X}-\underline{\mu})^T] = K_X$$

- K_X is symmetric, positive semi-definite matrix.
- Eigenvalue decomposition $K_X = U\Lambda U^T$.
- Let $A = U\sqrt{\Lambda}$, then $K_X = AA^T$, and

$$\underline{X} \overset{d}{=} A\underline{W} + \underline{\mu}$$

Consider another random vector $\underline{X}' = \sqrt{\Lambda}\underline{W}$, with independent entries $N(0, \lambda_i)$ distributed.

$$h(\underline{X}') = \sum_{i=1}^{n} h(X_i') = \frac{1}{2}\log(2\pi e)^n + \frac{1}{2}\sum \log \lambda_i$$

Now $h(X) = h(X') = h(\underline{W}) + \log\det(A)$.

- Can replace $\underline{W}$ by any other distribution
- **Important** Changing of coordinate system affects the differential entropy.

AEP

Theorem Let $X_1, \ldots, X_n$ be a sequence of i.i.d. r.v.'s with density $f(x)$.

$$-\frac{1}{n} \log f(X_1, \ldots, X_n) \to h(X)$$

in probability.

Definition typical set $A_\epsilon^{(n)}$:

$$A_\epsilon^{(n)} = \left\{ \underline{X}_1^n : \left| -\frac{1}{n} \log f(\underline{X}_1^n) - h(X) \right| \leq \epsilon \right\}$$

Theorem For any ϵ and large enough n

- $P(A_\epsilon^{(n)}) \geq 1 - \epsilon$
- $\text{Vol}(A_\epsilon^{(n)}) \leq 2^{n(H(X)+\epsilon)}$ for any n.
- $\text{Vol}(A_\epsilon^{(n)}) \geq 2^{n(H(X)-\epsilon)}$

Proof

$$\begin{aligned} 1 &= \int f(\underline{x}_1^n) d\underline{x}_1^n \quad \geq \int_{A_\epsilon^{(n)}} f(\underline{x}_1^n) d\underline{x}_1^n \\ &\geq 2^{-n(h(X)+\epsilon)} \int_{A_\epsilon^{(n)}} d\underline{x}_1^n \\ &= 2^{-n(h(X)+\epsilon)} \text{Vol}(A_\epsilon^{(n)}) \end{aligned}$$

Additive White Gaussian Noise Channel

Consider the channel

$$Y = X + W$$

with power constraint $E[X^2] \leq \sigma_X^2$, and $W \sim N(0, \sigma_W^2)$.

Definition

$$C = \max_{f_X : E[X^2] \leq P} I(X;Y)$$

Consider

$$\begin{aligned} I(X;Y) &= h(Y) - h(Y|X) \\ &= h(Y) - h(Y - X|X) \\ &= h(Y) - h(W) \\ &= h(Y) - \frac{1}{2} \log 2\pi e \sigma_W^2 \end{aligned}$$

$$E[Y^2] = E[X^2] + E[W^2] = \sigma_X^2 + \sigma_W^2$$

$$\begin{aligned} I(X;Y) &\leq \frac{1}{2} \log 2\pi e (\sigma_X^2 + \sigma_W^2) - \frac{1}{2} \log 2\pi e \sigma_W^2 \\ &= \frac{1}{2} \log \left(1 + \frac{\sigma_X^2}{\sigma_W^2} \right) \end{aligned}$$

Capacity as an Estimation Problem

Consider

$$\begin{aligned} I(X;Y) &= h(X) - h(X|Y) \\ &= h(X) - h(X - g(Y)|Y) \end{aligned}$$

for any function $g(.)$.

- Choose X to be $N(0, \sigma_X^2)$ distributed.
- Choose $g(.)$ to be the linear least square estimate of X. In the Gaussian case

$$\begin{aligned} g(Y) &= \frac{\sigma_X}{\sigma_X^2 + \sigma_W^2} Y \\ \text{var}[X - g(Y)] &= \frac{\sigma_W^2 \sigma_X^2}{\sigma_X^2 + \sigma_W^2} \end{aligned}$$

and $X - g(Y)$ is independent of Y. Now

$$\begin{aligned} I(X;Y) &= \frac{1}{2} \log(2\pi e \sigma_X^2) - \frac{1}{2} \log 2\pi e \frac{\sigma_W^2 \sigma_X^2}{\sigma_X^2 + \sigma_W^2} \\ &= \frac{1}{2} \log \left(1 + \frac{\sigma_X^2}{\sigma_W^2}\right) \end{aligned}$$

Discussions

- Denote $\hat{X} = g(Y)$, we call $\hat{X}$ a sufficient statistics if

$$X \to Y \to \hat{X}, X \to \hat{X} \to Y$$

- In Gaussian estimation problems (high dimension), the LLSE $\hat{X}$ satisfies this.
- $I(X;Y) = I(X;\hat{X})$. Processing Y to obtain a sufficient statistics does not reduce information.
- For general distributions of W with the same power, the LLSE $\hat{X}$, var$(X - \hat{X})$ is the same as the Gaussian case,

$$\begin{aligned} h(X - \hat{X}|Y) &\leq h(X - \hat{X}) \\ &\leq \frac{1}{2}\log 2\pi e \frac{\sigma_X^2 \sigma_W^2}{\sigma_X^2 + \sigma_W^2} \end{aligned}$$

- Equalities hold only for the Gaussian noise: **AWGN is the worst noise.**

A Mutual Information Game

- The transmitter tries to maximize the mutual information by choosing f_X, subject to a power constraint $E[X^2] = \sigma_X^2$.
- The channel (jammer) tries to minimize the mutual information by choosing a noise f_W, subject to a power constraint $E[W^2] = \sigma_W^2$.

Saddle point :

- the optimal input is Gaussian
- the worst noise is also Gaussian

More Realistic

Consider the channel

$$Y_i = X_i + W_i$$

where W_i is i.i.d. $N(0, \sigma_W^2)$, and the input has power constraint

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 \le \sigma_X^2$$

Theorem $C = \frac{1}{2}\log\left(1 + \frac{\sigma_X^2}{\sigma_W^2}\right)$ is the maximum achievable rate.

Proof outline:

- Generate random code book with 2^{nR} codewords, each of length n, with i.i.d. $N(0, \sigma_X^2 - \delta)$ entries.
- Joint typicality decoding.

To compute the error probability, w.o.l.g. assume the first codeword $\underline{x}(1)$ is transmitted.

- If the generated codeword violates the power constraint, claim an error.

$$E_0 = \left\{ \frac{1}{n} \sum_{i=1}^{n} x_i^2(1) \geq \sigma_X^2 \right\}$$

- Define

$$E_i = \{(\underline{X}(i), \underline{Y}) \text{ is jointly typical}\}$$

$$\begin{aligned} P(E_1) &\rightarrow 1 \\ P(E_i) &\approx 2^{-nI(X;Y)} \quad \text{for } i \neq 1 \end{aligned}$$

$$\begin{aligned} P_e^{(n)} &= P(E_0 \cup E_1^c \cup E_2 \ldots \cup E_{2^{nR}}) \\ &\leq \epsilon + \epsilon + 2^{nR} 2^{-n(I(X;Y)-\epsilon)} \end{aligned}$$

Converse

$$
\begin{aligned}
nR &= H(V) = I(V;\underline{Y}) + H(V|\underline{Y}^n) \\
&\le I(V;\underline{Y}) + 1 + nRP_e^{(n)} \\
&\le I(\underline{X};\underline{Y}) + 1 + nRP_e^{(n)} \\
&\le \sum_{i=1}^{n} I(X_i;Y_i) + 1 + nRP_e^{(n)}
\end{aligned}
$$

To drive $P_e^{(n)} \to 0$, need

$$R \le \frac{1}{n}\sum_{i=1}^{n} I(X_i;Y_i)$$

Key individual power constraint vs. average power constraint

Let $\frac{1}{n}\sum_i P_i \leq \sigma_X^2$.

$$\begin{aligned} R &\leq \frac{1}{n}\sum_{i=1}^{n} I(X_i;Y_i) \\ &\leq \frac{1}{n}\sum \frac{1}{2}\log\left(1+\frac{P_i}{\sigma_W^2}\right) \\ &\leq \frac{1}{2}\log\left(1+\frac{1}{n}\sum\frac{P_i}{\sigma_W^2}\right) \\ &= \frac{1}{2}\log\left(1+\frac{\sigma_X^2}{\sigma_W^2}\right) \end{aligned}$$

Corollary The concavity of the power-rate curve implies that we always want to spread the power evenly