LECTURE 16

Last time:

- Data compression
- Coding theorem
- Joint source and channel coding theorem
- Converse
- Robustness
- Brain teaser

Lecture outline

- Differential entropy
- Entropy rate and Burg's theorem
- AEP

Reading: Chapters 9, 11.

Continuous random variables

We consider continuous random variables with probability density functions (pdfs)

X has pdf $f_X(x)$

Cumulative distribution function (CDF)

$F_X(x) = P(X \leq x) = \int_{\infty}^{x} f_X(t)dt$

pdfs are not probabilities and may be greater than 1

in particular for a discrete Z

$P_{\alpha Z}(\alpha z) = P_Z(z)$

but for continuous X

$P(\alpha X \leq x) = P(X \leq \frac{x}{\alpha}) = F_X\left(\frac{x}{\alpha}\right) \int_{-\infty}^{\frac{x}{\alpha}} f_X(t)dt$

so $f_{\alpha X}(x) = \frac{dF_X(x)}{dx} = \frac{1}{\alpha} f_X\left(\frac{x}{\alpha}\right)$

Continuous random variables

In general, for $Y = g(X)$

Get CDF of Y: $F_Y(y) = \mathbf{P}(Y \le y)$ Differentiate to get

$$f_Y(y) = \frac{dF_Y}{dy}(y)$$

X: uniform on [0,2]

Find pdf of $Y = X^3$

Solution:

$$\begin{aligned} F_Y(y) &= \mathbf{P}(Y \le y) = \mathbf{P}(X^3 \le y) && (1) \\ &= \mathbf{P}(X \le y^{1/3}) = \frac{1}{2} y^{1/3} && (2) \end{aligned}$$

$$f_Y(y) = \frac{dF_Y}{dy}(y) = \frac{1}{6y^{2/3}}$$

Differential entropy

Differential entropy:

$$h(X) = \int_{-\infty}^{+\infty} f_X(x) \ln\left(\frac{1}{f_X(x)}\right) dx \qquad (3)$$

All definitions follow as before replacing P_X with f_X and summation with integration

$$\begin{aligned} & I(X;Y) \\ = & \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{X,Y}(x,y) \ln\left(\frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)}\right) dxdy \\ = & D\left(f_{X,Y}(x,y) || f_X(x)f_Y(y)\right) \\ = & h(Y) - h(Y|X) \\ = & h(X) - h(X|Y) \end{aligned}$$

Joint entropy is defined as

$$h(\underline{X}^n) = -\int f_{\underline{X}^n}(\underline{x}^n) ln\left(f_{\underline{X}^n}(\underline{x}^n)\right) dx_1 \ldots dx_n$$

Differential entropy

The chain rules still hold:

$h(X,Y) = h(X) + h(Y|X) = h(Y) + h(X|Y)$

$I((X,Y);Z) = I(X;Z) + I(Y;Z|Y)$

K-L distance $D(f_X(x)||f_Y(y)) = \int f_X(x) \ln\left(\frac{f_X(x)}{f_Y(y)}\right)$ still remains non-negative in all cases

Conditioning still reduces entropy, because differential entropy is concave in the input (Jensen's inequality)

Let $f(x) = -x\ln(x)$ then

$$\begin{aligned} f'(x) &= -x\frac{1}{x} - \ln(x) \\ &= -\ln(x) - 1 \end{aligned}$$

and

$$f''(x) = -\frac{1}{x} < 0$$

for $x > 0$.

Hence $I(X;Y) = h(Y) - h(Y|X) \geq 0$

Differential entropy

$H(X) \geq 0$ always

and $H(X) = 0$ for X a constant

Let us consider $h(X)$ for X constant

For X constant $f_X(x) = \delta(x)$

$$h(X) = \int_{-\infty}^{+\infty} f_X(x) \ln \left(\frac{1}{f_X(x)} \right) dx \qquad (4)$$

$h(X) \to -\infty$

Differential entropy is not always positive

See 9.3 for discussion of relation between discrete and differential entropy

Entropy under a transformation:

$h(X + c) = h(X)$

$h(\alpha X) = h(X) + ln\,(|\alpha|)$

Maximizing entropy

For $H(Z)$, the uniform distribution maximized entropy, yielding $\log(|\mathcal{Z}|)$

The only constraint we had then was that the inputs be selected from the set $\mathcal{Z}$

We now seek a $f_X(x)$ that maximizes $h(X)$ subject to some set of constraints

$f_X(x) \geq 0$

$\int f_X(x)dx = 1$

$\int f_X(x)r_i(x)dx = \alpha_i$ where $\{(r_1, \alpha_1), \ldots, (r_m, \alpha_m)\}$ is a set of constraints on X

Let us consider $f_X(x) = e^{\lambda_0 - 1 + \sum_{i=1}^{m} \lambda_i r_i(x)}$.
Let us show it achieves a maximum entropy

Maximizing entropy

Consider some other random variable Y with $f_y(y)$ pdf that satisfies the conditions but is not of the above form

$$
\begin{aligned}
h(Y) = & -\int f_Y(x)\,\mathrm{ln}(f_Y(x))dx \\
= & -\int f_Y(x)\,\mathrm{ln}\left(\frac{f_Y(x)}{f_X(x)} f_X(x)\right) dx \\
= & -D(f_Y||f_X) - \int f_Y(x)\,\mathrm{ln}(f_X(x))dx \\
\leq & -\int f_Y(x)\,\mathrm{ln}(f_X(x))dx \\
= & -\int f_Y(x)\left(\lambda_0 - 1 + \sum_{i=1}^{m} \lambda_i r_i(x)\right) dx \\
= & -\int f_X(x)\left(\lambda_0 - 1 + \sum_{i=1}^{m} \lambda_i r_i(x)\right) dx \\
= & \; h(X)
\end{aligned}
$$

Special case: for a given variance, a Gaussian distribution maximizes entropy

For $X \sim N(0, \sigma^2)$, $h(X) = \frac{1}{2}\mathrm{ln}(2\pi e \sigma^2)$

Entropy rate and Burg's theorem

The differential entropy rate of a stochastic process $\{X_i\}$ is defined to be $\lim_{n\to\infty} \frac{h(\underline{X}^n)}{n}$ if it exists

In the case of a stationary process, we can show that the differential entropy rate is $\lim_{n\to\infty} h(X_n|\underline{X}^{n-1})$

The maximum entropy rate stochastic process $\{X_i\}$ satisfying the constraints $E\left[X_i X_{i+k}\right] = \alpha_k$, $k = 0, 1, \ldots, p$, $\forall i$ is the p^{th} order Gauss-Markov process of the form

$$X_i = -\sum_{k=1}^{p} a_k X_{i-k} + \Xi_i$$

where the Ξ_is are IID $\sim N(0, \sigma^2)$, independent of past Xs and $a_1, a_2, \ldots, a_p, \sigma^2$ are chosen to satisfy the constraints

In particular, let $X_1, \ldots, X_n$ satisfy the constraints and let $Z_1, \ldots, Z_n$ be a Gaussian process with the same covariance matrix as $X_1, \ldots, X_n$. The entropy of $\underline{Z}^n$ is at least as great as that of $\underline{X}^n$.

Entropy rate and Burg's theorem

Facts about Gaussians:

- we can always find a Gaussian with any arbitrary autocorrelation function

- for two jointly Gaussian random variables $\underline{X}$ and $\underline{Y}$ with an arbitrary covariance, we can always express $\underline{Y} = \mathbf{A}\underline{X} + \underline{Z}$ for some matrix $\mathbf{A}$ and $\underline{Z}$ independent of $\underline{X}$

- if Y and X are jointly Gaussian random variables and $Y = X + Z$ then Z must also be

- a Gaussian random vector $\underline{X}^n$ has pdf

$$f_{\underline{X}^n}(\underline{x}^n) = \frac{1}{\left(\sqrt{2\pi}|\Lambda_{\underline{X}^n}|\right)^n}$$

$$e^{-\frac{1}{2}(\underline{x}^n - \underline{\mu}_{\underline{X}^n})^T \Lambda_{\underline{X}^n}^{-1} (\underline{x}^n - \underline{\mu}_{\underline{X}^n})}$$

where Λ and μ denote autocovariance and mean, respectively

- The entropy is $h(\underline{X}^n) = \frac{1}{2} \ln \left((2\pi e)^n |\Lambda_{\underline{X}^n}|\right)$

Entropy rate and Burg's theorem

The constraints $E\left[X_i X_{i+k}\right] = \alpha_k$, $k = 0, 1, \ldots, p$, $\forall i$ can be viewed as an autocorrelation constraint

By selecting the a_is according to the Yule-Walker equations, that give $p+1$ equations ion $p+1$ unknowns

$R(0) = -\sum_{k=1}^{p} a_k R(-k) + \sigma^2$

$R(l) = -\sum_{k=1}^{p} a_k R(l-k)$

(recall that $R(k) = R(-k)$) we can solve for $a_1, a_2, \ldots, a_p, \sigma^2$

What is the entropy rate?

$$\begin{aligned} h(\underline{X}^n) &= \sum_{i=1}^{n} h(X_i | \underline{X}^{i-1}) \\ &= \sum_{i=1}^{n} h(X_i | \underline{X}_{i-p}^{i-1}) \\ &= \sum_{i=1}^{n} h(\Xi_i) \end{aligned}$$

AEP

WLLN still holds:

$$-\frac{1}{n}\ln\left(f_{\underline{X}^n}(\underline{x}^n)\right) \to -E[\ln(f_X(x))] = h(X)$$

in probability for X_is IID

Define $Vol(A) = \int_A dx_1 \ldots dx_n$

Define the typical set $A_\epsilon^{(n)}$ as:

$$\left\{(x_1,\ldots,x_n) s.t. | -\frac{1}{n}\ln\left(f_{\underline{X}^n}(\underline{x}^n)\right) - h(X)| \leq \epsilon\right\}$$

By the WLLN, $P(A_\epsilon^{(n)}) > 1-\epsilon$ for n large enough

AEP

$$1 = \int f_{\underline{X}^n}(\underline{x}^n) dx_1 \ldots dx_n$$

$$1 \geq \int_{A_\epsilon^{(n)}} e^{-n(h(X)+\epsilon)} dx_1 \ldots dx_n$$

$$e^{n(h(X)+\epsilon)} \geq Vol(A_\epsilon^{(n)})$$

For n large enough, $P(A_\epsilon^{(n)}) > 1 - \epsilon$ so

$$1 - \epsilon \leq \int_{A_\epsilon^{(n)}} f_{\underline{X}^n}(\underline{x}^n) dx_1 \ldots dx_n$$

$$1 - \epsilon \leq \int_{A_\epsilon^{(n)}} e^{-n(h(X)-\epsilon)} dx_1 \ldots dx_n$$

$$1 - \epsilon \leq Vol(A_\epsilon^{(n)}) e^{-n(h(X)-\epsilon)}$$

6.441 Information Theory
Spring 2010