| | |
|---|---|
| **6.441 Transmission of Information** | April 11, 2006 |
| Lecture 15 | |
| *Lecturer: Madhu Sudan* | *Scribe: Brandon Roy* |

# Today

- Differential entropy

  Conditional entropy, Joint entropy, Mututal information...

- Channel capacity

# Admin

- PS3 due tomorrow
- Office hours, Thursday afternoon (send email)

# Motivations from last time

Recall the "6.441 channel". We had input $X \in [-1, 1]$, noise $W \sim \text{Uniform}[-\epsilon, \epsilon]$, and output $Y = X + W$. We saw that

- If $\epsilon = 0$, channel has infinite capacity.
- If $\epsilon > 0$, channel has finite capacity.

### Differential Entropy

Beginning with differential entropy, introduced last time, let us analyze this channel. We have $X$ taking values in $\mathbb{R}$ with pdf $f = f_X$. Recall that we are working with $X_\epsilon$, the $\epsilon$-discretization of $X$. Then

$$h(X) \triangleq \lim_{\epsilon \to 0} \{H(X_\epsilon) + \log \epsilon\} = -\int_{-\infty}^{\infty} f_X(x) \log(f_X(x))dx \quad \text{(if well behaved)}$$

Differential entropy is similar to "discrete" entropy but it is important not to draw too many conclusions from this similarity. For example, consider the following:

- $X \sim \text{Uniform}(a, b)$
- $h(X) = \log(b - a)$

- $h(aX) = h(X) + \log|a|$

Note that for some choices of $a$, goes to $\infty$, or if $b-a$ is very small, $\log(b-a) < 0$. So caution: $\exists X$ s.t. $h(X) < 0$ which is never true with $H(X)$ (when $X$ is discrete)

## Definitions

We now proceed to develop concepts for continuous random variables along the lines of those developed for discrete random variables. Consider a collection of random variables $X_1 \dots X_n$ (real-valued) with pdf $f(X_1, \dots, X_n)$.

**Joint Entropy**

$$h(X_1, \dots, X_n) = -\int_{X_1,\dots,X_n} f(x_1, \dots, x_n) \log f(x_1, \dots, x_n) dx_1 \dots dx_n$$

**Conditional Entropy**

Consider $(X, Y)$ with joint distribution $f(X, Y)$, marginal distributions $f_X$,$f_Y$, and conditional distribution $f_{X|Y}(x|y)$. Then

$$\begin{aligned} h(X|Y) &= -\int_Y f_Y(y) \left[ \int_X f_{X|Y}(x|y) \log f_{X|Y}(x|y) dx \right] dy \\ &= -\int \int_{X,Y} f(x, y) \log f_{X|Y}(x|y) dx dy \end{aligned}$$

**Divergence**

The divergence between pdf's $f$ and $g$ is

$$D(f||g) = \int_X f(x) \log \frac{f(x)}{g(x)} dx$$

Furthermore,

$$D(f||g) \geq 0 \qquad \text{(usual proof by Jensen's Inequality)}$$

Applying this,

$$(x, y) : D(f||f_X, f_Y) \geq 0 \implies h(X|Y) \leq h(X)$$

(Conditioning reduces entropy)
Note: when *comparing* entropies, any "$\log \epsilon$" terms show up on both sides and the comparison makes sense. Generally however, this is not true for the actual "values".

**Mutual Information**

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

If $X$ and $Y$ are "continuations" (opposite of discretizations) of discrete $\tilde{X}$, $\tilde{Y}$ then $I(X;Y) = I(\tilde{X}; \tilde{Y})$.

**Chain Rule**

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

# Maximum entropy distributions

## Uniform distribution

Among random variables $X$ taking values in $[0, 1]$ the differential entropy is maximized by the $X \sim \text{Uniform}(0, 1)$.

**Proof 1**

Let $X$ be any r.v. taking values in $[0, 1]$.
Let $Y$ be any r.v. with distribution Uniform$(0, 1)$, independent of $X$.
Let $Z = (X + Y) \mod 1$

Then
  $f_Z$ is Uniform$(0, 1)$ (not hard to show)
  $f_{Z|X}$ is Uniform$(0, 1)$

$$\begin{aligned} h(Y, Z) &= h(X, Y) \\ &= h(X) + h(Y) \end{aligned}$$

$$h(Y, Z) \leq h(Y) + h(Z)$$

$$\implies h(X) \leq h(Z)$$

**Proof 2 (Chung's proof)**

$$\begin{aligned} h(X) &= E\left[\log \frac{1}{p(X)}\right] \\ &\leq \log \left[E \frac{1}{p(X)}\right] && \text{(Jensen's inequality)} \\ &= \log \left[\int_S p(x) \frac{1}{p(x)} dx\right] && \text{($S$ is the support set)} \\ &= \log |X| \end{aligned}$$

which is the entropy of the uniform distribution.

So to conclude, among random variables taking values in $[0, 1]$ the differential entropy is maximized by $X \sim \text{Uniform}(0, 1)$.

## Gaussian distribution

Furthermore, among (unbounded) random variables with mean 0 and variance 1, the differential entropy is maximized by $X \sim \text{Normal}(0, 1)$. In other words, for any
  $X'$ distributed arbitrarily with mean 0 and variance 1
  $X \sim \text{Normal}(0, 1)$

$$D(X' || X) = h(X) - h(X') \geq 0$$

The Gaussian distribution has maximum entropy.

**Entropy of the Gaussian distribution**

Let $X \sim \text{Normal}(0, \sigma^2)$. Denote the pdf of $X$ by $\Phi(X)$ Note that $\log \Phi(x) = a + bx^2$. Then

$$\begin{aligned} h(X) &= -\int \Phi(x) \log \Phi(x) dx \\ &= a \int \Phi(x) dx + b \int x^2 \Phi(x) dx \\ &= a + b\sigma^2 \end{aligned}$$

# AEP Theorem

If $X_1, \ldots, X_n$ iid. $\quad X$ then

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

in probability

**Typical set**

$$A_\epsilon^{(n)} = \left\{ (x_1, \ldots, x_n) : \left| -\frac{1}{n} \log f(x_1, \ldots, x_n) - h(X) \right| \leq \epsilon \right\}$$

Also, define the "volume" of a set $S$ as

$$Vol(S) = \int 1_S dx_1 \ldots dx_n$$

Then, $\forall \delta, \epsilon > 0$, $\exists n_0$ s.t. $\forall n \geq n_0$:

1. $\Pr(A_\epsilon^{(n)}) \geq 1 - \delta$
2. $Vol(A_\epsilon^{(n)}) \leq 2^{(h(X)+\epsilon)n}$
3. $Vol(A_\epsilon^{(n)}) \geq (1 - \delta) 2^{(h(X)-\epsilon)n}$

**Proofs**

**1:**
$\Pr(A_\epsilon^{(n)}) \geq 1 - \delta$. Follows from the LLN, applied to continuous random variables.

**2:**

$$\begin{aligned} 1 &= \int f(x_1, \ldots, x_n) dx_1, \ldots, dx_n \\ &\geq \int 1_{A_\epsilon^{(n)}} f(x_1, \ldots, x_n) dx_1, \ldots, dx_n \\ &\geq \int 1_{A_\epsilon^{(n)}} 2^{-(h(X)+\epsilon)n} dx_1, \ldots, dx_n \\ &= 2^{-(h(X)+\epsilon)n} \cdot Vol(A_\epsilon^{(n)}) \end{aligned}$$

$$\implies Vol(A_\epsilon^{(n)}) \leq 2^{(h(X)+\epsilon)n}$$

**3:**

$$\begin{aligned}1-\delta &\le \int 1_{A_\epsilon^{(n)}} f(x_1,\ldots,x_n)dx_1,\ldots,dx_n \\ &\le \int 1_{A_\epsilon^{(n)}} 2^{-(h(X)-\epsilon)n} dx_1,\ldots,dx_n\end{aligned}$$

$$\implies Vol(A_\epsilon^{(n)}) \ge (1-\delta)2^{(h(X)-\epsilon)n}$$

# Channel capacity

Now, back to the beginning. Recall our "6.441 channel": $Y = X + W$. Suppose $2\epsilon = \frac{1}{k}, k \in \mathbb{Z}$. We expected the "intuitive capacity" $\ge \log\lfloor 1 + \frac{2}{2\epsilon}\rfloor$.

## Capacity

Define capacity as

$$C = \max_{f_X} \{I(X;Y)\}$$

Note that the maximization is over all distributions subject to constraints. But this is just a definition, let's see if it makes sense for our channel.

$$\begin{aligned}\max_{f_X}\{I(X;Y)\} &= \max_{f_X}\{h(Y) - h(Y|X)\} \\ &= \max_{f_X}\{h(Y) - h(X+W|X)\} \\ &= \max_{f_X}\{h(Y) - h(W|X)\} \\ &= \max_{f_X}\{h(Y) - h(W)\} \\ &\le \log(2(1+\epsilon)) - \log(2\epsilon) \\ &= \log\left(\frac{1}{\epsilon}+1\right)\end{aligned}$$

Wish to prove: operational capacity $\le$ formal capacity. "Converse coding theorems" We want to find upper bound on $R$. The sequence of actions in transmission is

Choose $\underline{x} = (x_1,\ldots,x_n) \in$ set $M$ of size $2^{nR}$
Receiver gets $\underline{y} = (y_1,\ldots,y_n)$
We guess $\underline{\hat{x}} = (\hat{x_1},\ldots,\hat{x_n})$.

So we have the Markov chain $\underline{X} \to \underline{Y} \to \underline{\hat{X}}$ and use Fano's Inequality: $H(X|Y) \le 1 + P_e \log|M|$

$$\begin{aligned}I(X;Y) &= H(X) - H(X|Y) \\ &\ge H(X) - (1+\log|M|P_e) \\ &\ge \log|M|(1-P_e) - 1 \\ &= nR(1-P_e) - 1\end{aligned}$$

Note that above, we are using "discrete entropy" since $X$ is "$\epsilon$-discretized"

But we also have

$$I(X;Y) = h(Y) - h(Y|X) \leq \sum_{i=1}^{n} h(y_i) - h(y_i|x_i) = \sum_{i=1}^{n} I(x_i;y_i)$$
$$\leq nC$$

and combining these two inequalities, we have

$$R \leq C$$