ECE 287 Apr. 11

\[ I(X; Y) = D\left( p(x, y) \parallel p(x)p(y) \right) \]
\[ = \min_{q(y)} \sum_{x} p(x) D\left( q(y|x) \parallel q(y) \right) \]
\[ \text{min attained by } p(y) = \sum_{x} p(x) p(y|x) \]

Information Capacity

\[ C = \max_{p(x)} I(X; Y) \]

\[ = \max_{p(x)} \min_{q(y)} \sum_{x} p(x) D\left( q(y|x) \parallel q(y) \right) \]
\[ = \min_{q(y)} \max_{p(x)} D\left( p(y|x) \parallel q(y) \right) \]

Furthermore, if \( p(x) \) attains the maximum in (1),
then \( q(y|x) = \frac{p(x)}{p(y|x)} \) attains the minimum in (2) and (3).

Differential Entropy

Let \( X \) be a continuous RV with probability density function \( p_X(x) \).

\[ \frac{dP}{dx}(X \leq x) = \begin{cases} p_X(x) & \text{for } x \neq a \\ 0 & \text{for } x = a \end{cases} \]

Then the differential entropy of \( X \) is \( h(X) = \int p_X(x) \log \frac{1}{p_X(x)} dx \)

\[ = -\int p_X(x) \log \frac{1}{p_X(x)} dx \]

Example

1) If \( X \) is uniform \( X \sim \text{unif}[a, b] \), then \( h(X) = \log (b-a) \)

2) If \( X \sim \text{N}(\mu, \sigma^2) \) then \( h(X) = \frac{1}{2} \log (2\pi e \sigma^2) \), then \( h(X) = \frac{1}{2} \log (2\pi e \sigma^2) \)
Properties

1. \( h(x) \) is invariant in coordinate shift. i.e., \( h(x) = h(x + a) \)
2. \( h(x) \) as a function of the pdf \( p(x) \) is concave
3. \( h(ax) = \log |a| \) if \( a \neq 0 \)
4. \( h(x) \) can be negative, unlike discrete

Let \( Y \) with \( X = x \), then \( p(y|x) \) be continuous with pdf \( p(y|x) \)
Then the conditional differential entropy is

\[
h(Y|X) = \int p(y|x) \log p(y|x) \, dy \quad \text{if } X \text{ is discrete}
\]
\[
h(Y|X) = \int p(y|x) \log p(y|x) \, dy \quad \text{if } X \text{ is continuous}
\]

By concavity and Jensen's inequality

\[
h(Y|X) \leq h(Y) \quad \text{iff } X \text{ and } Y \text{ are independent}
\]

when \( (X,Y) \) are jointly continuous, the joint entropy is

\[
h(X,Y) = E[\log p(x,y)] = h(X) + h(Y|X) = h(Y) + h(X|Y)
\]

Chain rule

\[
h(X^n) = h(X_1) + h(X_2 | X_1) + \cdots + h(X_n | X^{n-1}) = \sum_{i=1}^{n} h(X_i)
\]

Ex.

If \( X \sim N(\mu, k) \), then

\[
h(X^n) = \frac{1}{2} \log \left( 2\pi e n \sigma_k^2 \right) = \sum_{i=1}^{n} \frac{1}{2} \log \left( 2\pi e k_{ii} \right)
\]

\[ |k| = \prod_{i=1}^{n} k_{ii} \quad \text{Hadamard's inequality} \]
Let \( p(x) \) and \( q(x) \) be two pdfs on \( \mathbb{R} \), then the relative entropy is
\[
D(p||q) = \int p(x) \log \left( \frac{p(x)}{q(x)} \right) dx
\]

Let \( X \) and \( Y \) be jointly continuous with joint pdf \( p(x,y) \)

Then
\[
I(X;Y) = \mathbb{E} \left[ \log \frac{p(x,y)}{p(x)p(y)} \right] = h(X) + h(Y) - h(X,Y) = h(Y) - h(Y|X)
\]

Let \( X \sim N(0,p) \) and \( Z \sim N(0,N) \) be independent.
Let \( Y = X + Z \sim N(0, p+N) \) then
\[
I(X;Y) = h(Y) = h(Y|X)
\]
\[
= \frac{1}{2} \log (2\pi e (p+N)) = \frac{1}{2} \log (2\pi e N)
\]
\[
= \frac{1}{2} \left( \log \left( \frac{p+N}{N} \right) = \frac{1}{2} \log \left( 1 + \frac{p}{N} \right) \right)
\]

Let \( X \) be discrete with pmf \( p(x) \) and \( Y \mid X = x \) be continuous with conditional pdf \( p(y|x) \). Let \( x \)

- Then the marginal pmf of \( Y \) is
\[
p(y) = \sum_x p(x)p(y|x)
\]
and the conditional pmf of \( X \) given \( Y \)
\[
p(x|y) = \frac{p(x)p(y|x)}{p(y)} = \frac{P(x)p(x|y)}{\sum_x p(x)p(y|x)} \quad \text{Bayes' Rule}
\]

- The mutual information between \( X \) and \( Y \) can be written as
\[
I(X;Y) = H(X) - H(X|Y) = h(Y) - h(Y|X)
\]
General minimax theorem for information capacity

Let \( \{p(y|x)\}_{x \in X} \) be an arbitrary collection of probability distribution on \( Y \).

**Minimax theorem (Kempf)\(^1\)**

\[
\sup_{p(x)} \inf_{Q(y|x)} \int D(p(y|x)||Q(y|x)) \, dQ(y|x)
\]

For any finite probability measure on \( X \),

\[
\inf_{p(x)} \sup_{Q(y|x)} D(p(y|x)||Q(y|x))
\]

Let \( X = \{X_n\}_{n=0}^{\infty} \) be a random process in a finite alphabet \( X \).

Then its entropy rate is

\[
\overline{H}(x) = \lim_{n \to \infty} \frac{1}{n} H(X_0, \ldots, X_n) \text{ if the limit exists.}
\]

Ex

\[1 \text{ If } X \text{ is iid, } \overline{H}(x) = H(X_0).\]

\[2 \text{ If } X \text{ is aperiodic irreducible Markov chain with unique stationary distribution, then } \overline{H}(x) = \lim_{n \to \infty} H(X_0 | X_n), \quad \text{time invariant}\]

\[3 \text{ Let } Y = \{Y_n\}_{n=0}^{\infty} \text{ be a stationary Markov chain and } X_n = f(Y_n).\]

Then \( X = \{X_n\}_{n=0}^{\infty} \) is hidden Markov

and \( H(X_n | X^n, Y) \leq H(X) \leq H(X_n | X^n) \)

Same limit as \( n \to \infty \).
(a) Let $X = (X_n)_{n=0}^\infty$ be stationary.

Then $H(X) = \lim_{n \to \infty} \frac{1}{n} H(X_1, \ldots, X_n) = \lim_{n \to \infty} \frac{1}{n} H(X_n | X_{n-1})$

$\Rightarrow \frac{1}{n} \sum_{k=1}^{n} a_k \to a$

Ergodicity

Let $X = (X_n)_{n=0}^\infty$ be a random process
Let $\Omega$ be the associated sample space, namely,

$\Omega = \mathcal{X}^\mathbb{Z} = \{ \ldots, X_n, X_0, X_1, \ldots \}$,

and

$X_n(\omega) = X_n$

$n$th coordinate of the function

Let $T$ be a time shift operator on $\Omega$, that is,

$\Omega = \{ \ldots, W_1, W_0, W'_1, \ldots \}$

$T$-center

Then $T w = (\ldots, w, W_1, w, \ldots)$ shifted

A stationary process satisfies (by definition)

$P(T^{-1} A) = P(w) \cdot P(T w \epsilon A) = P(A)$ every measurable $A$.

We say that an event $A$ is shift-invariant if

$A = T^{-1} A$

\[ T \begin{pmatrix} 0.1 \end{pmatrix} = \begin{pmatrix} 0.1 \end{pmatrix} \]

\[ A = \{ 0 \ldots, 010101 \ldots, \ldots, 100101 \ldots \} \text{ is shift-invariant} \]
We say that \( X = (X_n)_{n \to \infty} \) is ergodic if \( A = T^*A \) implies \( \text{p}(A) = 0,1 \).